UNDERSTANDING AND CHARACTERIZING ENGINEERING RESEARCH DATA FOR ITS BETTER MANAGEMENT

TOM HOWARD, MANSUR DARLINGTON, ALEX BALL, CHRIS MCMAHON AND STEVE CULLEY

erim2rep100420mjd10.pdf

ISSUE DATE: 17 September 2010

Catalogue Entry

Title	Understanding and Characterizing Engineering Research Data for its Better Management
Creator	Tom Howard, Mansur Darlington (authors)
Contributor	Alex Ball, Chris McMahon, Steve Culley
Subject	Data preparation; data development; data records; data description; ERIM Data Management Terminology; Research Activity Information Development modelling
Description	Engineering research data is diverse in character, spanning everything from material properties to questionnaire responses and interview transcripts. As a first step towards improving its management, a terminology is being developed to help describe the different types of engineering information and the different forms of development process and management activity to which it is subjected to during research. In particular the notions of 'data purposing', 'data re-purposing' and 'supporting data re-use' have been identified as data preparation activities which motivate research data management. This terminology has evolved in tandem with a scoping survey and audit of selected cases of engineering research data, and underlies a new modelling method for visualizing the associations between research data objects. These Research Activity Information Development (RAID) diagrams support both the initial researchers in managing the data and later re-users in understanding the data.
Publisher	University of Bath
Date	20th April 2010 (creation)
Version	1.0
Type	Text
Format	Portable Document Format (PDF/A-1b:2005)
Resource Identifier	erim2rep100420mjd10
Language	English
Rights	

Citation Guidelines

Tom Howard, Mansur Darlington, Alex Ball, Chris McMahon and Steve Culley. (2010). *Understanding and Characterizing Engineering Research Data for its Better Management* (version 1.0). ERIM Project Document erim2rep100420mjd10. Bath, UK: University of Bath.

1. INTRODUCTION

Motivated by the general drive toward accountability and efficiency in public sector activities, there is a developing interest in the ways that the data gathered and generated in publically-funded research might be made more available for use by the research community at large.

In addition to this, it is recognised that researchers themselves can benefit from and wish to have easier access to existing data (Beagrie, et al. 2009) yet, because of poor management, social and commercial pressures and legacy sharing practice, such access is often not possible (Birnholtz & Beitz, 2003).

Nevertheless, the practice of sharing, and management for sharing, is more widespread in some disciplines than others (e.g. the space sciences, bio-research) as a result of necessity and technical opportunity; and because of the demands of funders (Jones, 2009). Many aspects of the management of research data are considered under the general heading of 'curation', support for which is provided by the work done, for example, by the DCC (DCC, 2007). Much of this is helpful at a general level in assisting researchers plan and implement better management of research data for their own and then future use.

The work reported here builds on the curation work done by others in a specialized and more detailed way. It attempts to identify and characterize in particular the spectrum of research data that is characteristically collected and generated during engineering research activities. The purpose of this study is to provide a sounder basis for the management of such data so that it can be made more amenable to use, re-use and re-purposing (see definitions below) by which means the value of the data is increased. From greater sharing of data would be derived a number of benefits including the reduction in duplicated work, greater transparency of research, an improved basis for validation and a lesser need for collection and generation (Birnholtz & Beitz, 2003; Fry, et al., 2008).

Although each researcher will have an understanding of their own research data, little is known about the diversity, character, creation, collection and use of engineering research data as it exists across the engineering research community. Even less is known about how this data is currently managed and how it might be managed better.

The authors' investigations consist of two main parts. The first part consists of the identification and characterization of engineering research data through the inspection of data generated and collected in two repositories which constitute a representative set of engineering research data. One of these is the KIM Grand Challenge Project (Darlington, et al., 2009) and the other is the repository of research data held by the I*d*MRC at the University of Bath (http://www.bath.ac.uk/idmrc/). The second has been the subject of a data audit using the Digital Curation Centre's Data Audit Framework (Jones, et al., 2008). Both repositories contain research data generated by a diverse set of research studies broadly within the engineering domain. From this, by inspection and by partitioning the data along a number of selected dimensions, a sub-set of candidate research data has been identified which might reasonably be considered to represent the full spectrum of engineering research data. Each instance of this is the data related to a particular research activity or project. For each project or activity one or more examples of data will have been generated. The scrutiny of the data itself is augmented by interviews with the researchers involved in the project and the data

gathering/generation. This not only provides the basis for an understanding of the data itself and the processes to which it is subject during the research activity, but also of the management practices (or lack thereof) to which the data has been subject.

The second part of the work is theoretical, and is the basis for the analysis of the subject research data in respect of its management. Understanding how to manage research data requires not only that the nature of the data is understood, but so too the context in which the data is collected, gathered, manipulated and used. In short, to understand the management of data it is necessary to understand the data life cycle at appropriate levels of analysis and its influence on the data. It is the theoretical aspects that are dealt with first here, in Section 3. Readers are directed also to a separate work carried out as part of the ERIM project (Ball, 2010) which is a recent review of the literature on research (and other) data curation. Following laying the theoretical ground-work, the empirical work is presented in Sections 4-6.

2. LITERATURE REVIEW

This literature review will contain a brief summary of some research data lifecycles. There has been a number of characterizations of the data life cycle developed for a number of different purposes (see Ball, 2010 for a concise overview). Of particular interest is the high-level life cycle model adopted by the Digital Curation Centre (Section 2.1) and in contrast the research activity model proposed by the I2S2 project (Section 2.2).

2.1 DCC's Curation Lifecycle Model

The DCC's model has been produced for aligning curation tasks with the life cycle stages of a digital object, intended as a planning tool for data creators, curators and users (Higgins, 2008). A graphical representation of the model can be seen as Figure 2-1.

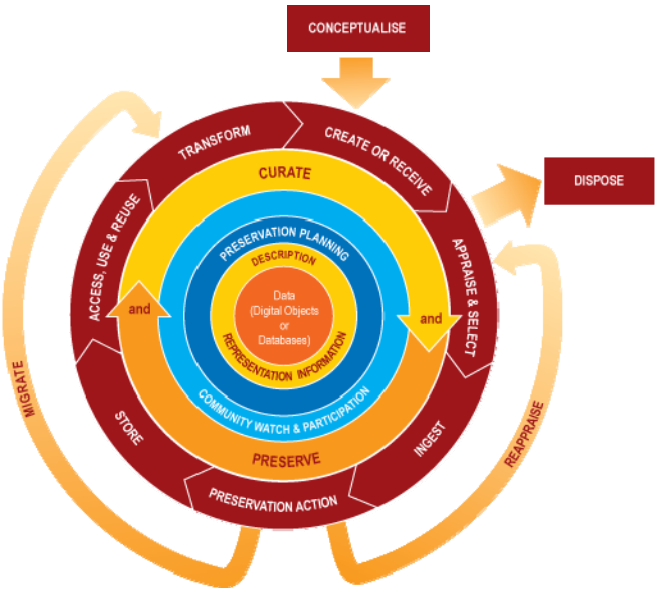

Figure 2-1. The DCC's Curation Lifecycle Model

At the centre of the Model is the digital data, which is here identified with simple and complex digital objects or databases. The model notes three levels of full-life-cycle actions:

- *Description* and (management of) *Representation Information*. The creation, collection, preservation and maintenance of sufficient metadata (and recursions thereof) to enable the data to be used and re-used for as long as it has value to justify continued curation.
- *Preservation Planning*. Strategies, policies and procedures for all curation actions.
- *Community Watch and Participation*. The observation of the target community of the data, in order to track changes in their requirements for the data, and participation in the development of standards, tools and software relevant for the data.

The fourth level, *Curate and Preserve*, properly describes most of the actions in the model, but is used here to represent the execution of the planned management and administrative actions supporting curation.

Notwithstanding its usefulness as a basis for planning curation and the identification of some key concepts, this model – because of the high-level description – is unsuited for use in the analysis of data in respect of its management.

2.2 I2S2 Research Activity Life Cycle Model

A second model of the data life cycle (Figure 2-2) has been developed in the I2S2 (Infrastructure for Integration in Structural Sciences) project. Whereas the DCC model is intended to support the 'curators' of data, the I2S2 model recognizes the – often quite different – needs of researchers in supporting data re-use.

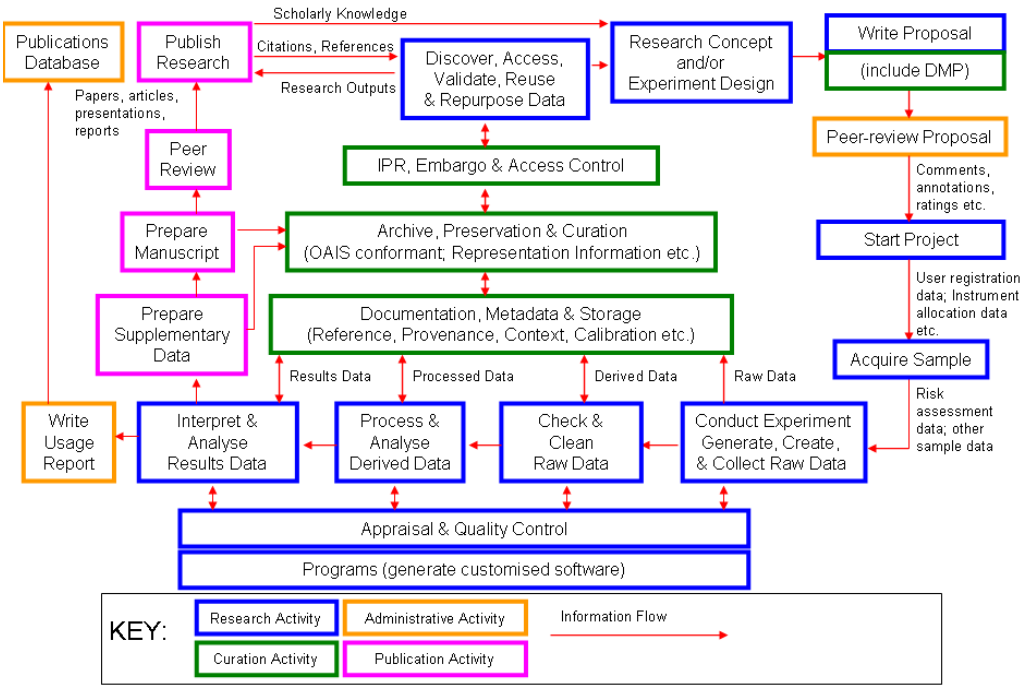

Figure 2-2. The I2S2 idealised Research Activity Life Cycle Model

The I2S2 model represents the processes and phases of a typical science experiment project. As indicated in the blue text boxes, the stages include: development of the research proposal; its peer-review; carrying out of the experiment; and processing, analysis and interpretation of the data which is eventually reported and published in various forms as research outputs. In addition the model incorporates a number of idealised stages (green boxes) to cater for the long-term management and availability of scientific data; these include: appraisal and quality control; documentation including metadata and contextual information; storage, archive, preservation and curation; and IPR, embargo and access control.

2.3 Overview

Both models identify and introduce some important concepts in relation to data use and management; however, for the purposes of analysis what is needed is a model which identifies in some detail and defines the sub-activities or processes underpinning those identified in the DCC and I2S2 models, which result in the creation and mutation of data as a result of the research activity itself. Without understanding the processes that data undergoes at this lower level a full understanding of the need for management intervention at critical points in the life cycle cannot be achieved.

The authors have developed a model which identifies these more basic sub-activities and processes and the lexicon with which these activities can be discussed. The model is validated by the work reported in Section 4 and those following.

3. THEORY AND TERMINOLOGY

In this section we discuss the theory and terminology associated with the different data management activities and propose a means by which the relationships between data records and the development processes that lead to data change can be modelled. This approach will then be developed to allow the development of data in individual research projects or activities to be mapped at the level of granularity necessary to ensure good management (see Section 5).

3.1 Modelling the Research Activity in Respect of the Data Life Cycle

The terms *use, re-use* and *re-purposing* have been introduced above. In order to be able to communicate about the data life cycle and the sub-activities and processes that go on during research, these terms, together with others, require definition. Indeed, the authors have found that their research, their thinking and, indeed, their intra- and extra-team communication has been hampered by the lack of a clear vocabulary (shared or otherwise) with which to discuss the territory being mapped, and no complete existing lexicon associated with this work has been found. This limitation motivated the adoption of a set of prescriptive terms, some, but not all, borrowed from existing work (see Appendix A for the developing Data Management Terminology and notes on provenance) which will be introduced in the following sections, including the five activities defined as follows:

- **Data Use** Using research data for the current research purpose/activity to infer new knowledge about the research subject.

- **Data Re-use:** Using research data for a research purpose/activity other than that for which it was intended.
- **Data Purposing** Making research data available and fit for the *current research activity*.
- **Data Re-purposing** Making existing research data available and fit for a *future **known** research activity*
- **Supporting Data Re-use**: Managing existing research data such that it will be available for a *future **unknown** research activity.*

  Regrettably it has not been possible to coin an apposite verb for this activity; however it seems likely that 'supporting data re-use' consists of some or all of the activities associated with archiving, preservation and curation.

The first two items concern 'use' of data. The last three items concern the preparation of or readying of data for use, which will be referred to as 'preparation activities'. There is an important relationship between data development, management and use that runs throughout research.

Logically associated with each of the preparation activities is a number of data development processes that the data undergoes and through which, in general, this data is organized, mutated and multiplied. When data, data records or cases are (as appropriate) subjected to any of these 'preparation activities' and 'development processes' then it can be said that management is being carried out.

In Table 3-1 the authors show what they believe to be the logical relationship between the identified preparation activities and the development activities (see Appendix A for formal definitions and their provenance):

Table 3-1. Data development and the three preparation activities

	Data Preparation for		
Data Development	***Purposing***	***Re-purposing***	***Supporting Re-use***
Addition	✓	×	×
Association	✓	✓	✓
Aggregation	✓	×	×
Annotation	✓	✓	✓
Augmentation	✓	✓	×
Collection	✓	×	×
Collation	✓	✓	✓
Deletion	×	✓	×
Derivation	✓	×	×
Duplication	×	×	×
Extraction	✓	✓	×
Generation	✓	×	×
Migration	✓	✓	×
Population	✓	×	×
Refinement	✓	✓	×

Some validation of these processes has been found in the data audit cases presented in Section 7.1. There may be other legitimate data development processes that can be added to this list. In addition to having logical connections with the sub-activities, the three data preparation activities can be partitioned in terms of when they occur and by whom and for whom they are carried out. Purposing is an activity carried out by a researcher on their own data for use in the current research activity. In contrast to this use of data now, both supporting data re-use and re-purposing are carried out for future activities, in the first instance on data for which the future use is unknown, in the second on data the future use of which is already known.

Clearly the second condition allows more extensive and directed preparation of the data than does the first, and therefore implicates a greater number of the development processes. These differing conditions applying to the three preparation activities are reflected in Figure 3-1. Their typical relation to research project duration is suggested in Figure 3-2.

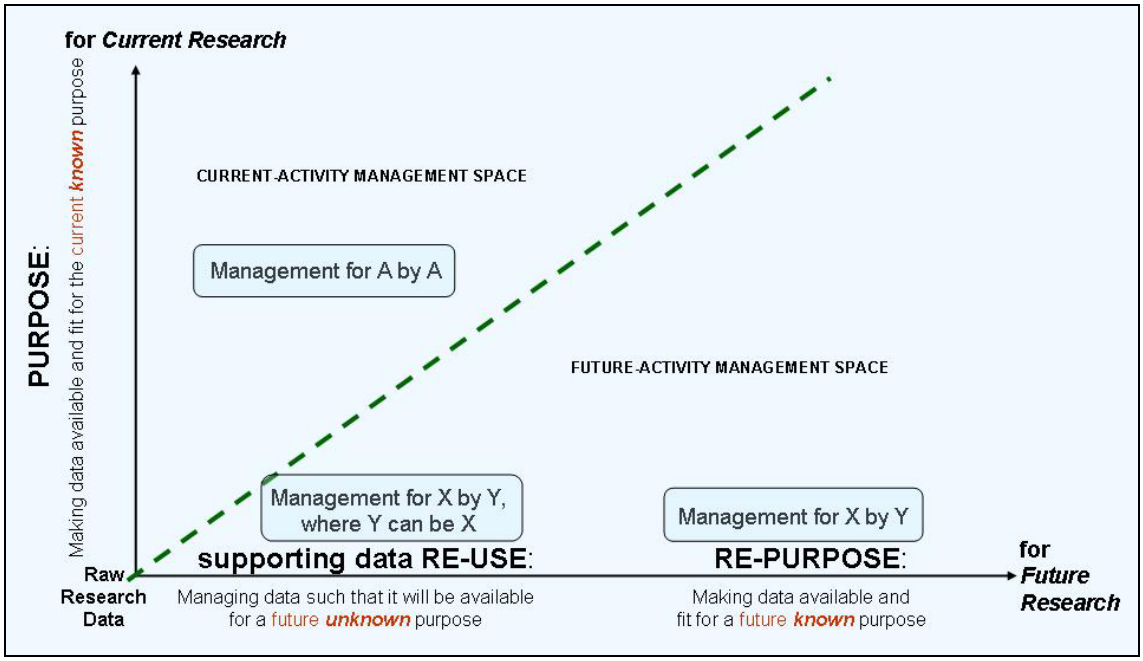

Figure 3-1. The relation between the three data preparation (management) tasks.

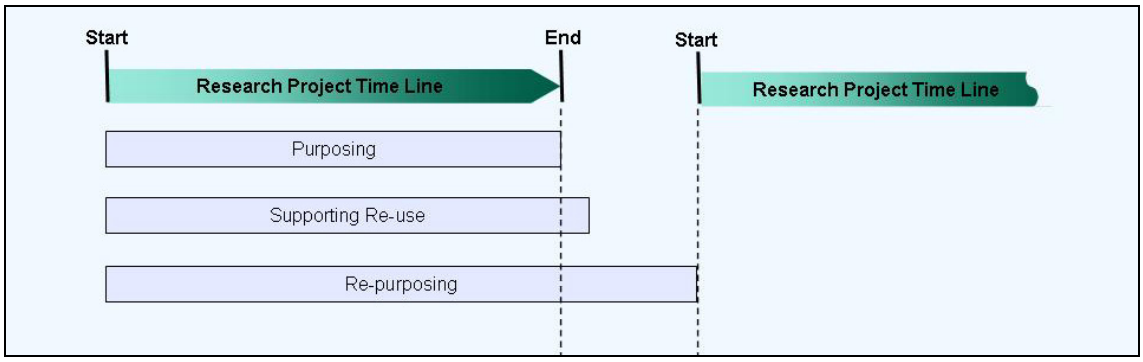

Figure 3-2. The data preparation activities and the research project time line

3.2 Managing Objects, Relationships and Levels

Scrutiny of engineering research data suggested to the authors that there are two further important dimensions which must be understood for the purposes of better data management. The first of these concerns precisely what sort of data 'objects' might be the subject of management and the second concerns the level of granularity of management.

Clearly, it is possible to manage data (that is to say, subject it to the sort of processes identified above) at the level of the data itself. However, data usually comes in what can be thought of as 'containers'. In electronic form these containers consist of discrete files, either single or in associated groups. Similarly, data can be manifest in physical form as, for example, a sheet of paper or a bound volume. In fact, data can exist in a number of forms and levels of granularity (Darlington, et al., 2008).

In their research work the authors have been using the term 'data asset' to refer to any object which might provide 'information associated with the research activity'. This includes the physical objects, including both data objects and physical things such as lab samples, which in a sense constitute 'data'. The term data asset likewise embraces all *digital* objects associated with research. Observation suggests that there are a number of distinct digital objects that may be of interest which perform different rôles in the research process and which could have different management requirements.

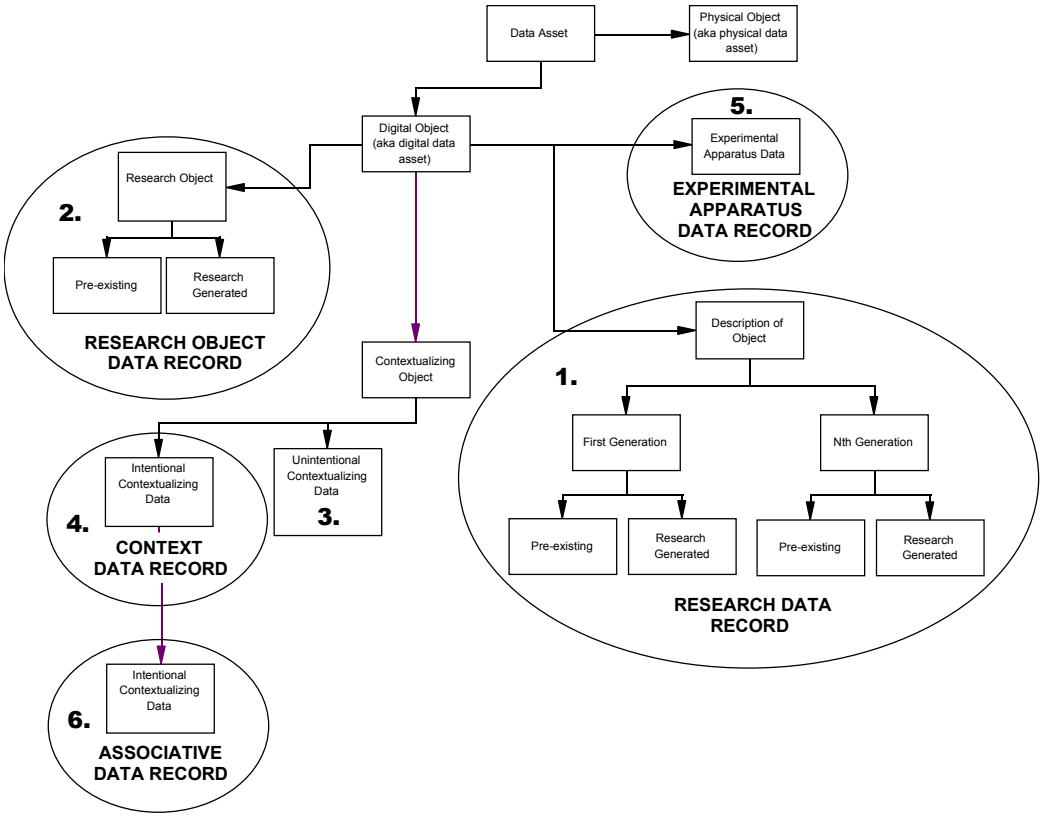

Figure 3-3. The Research Activity Data Object Taxonomy (the numbering is for reference purposes only)

The digital objects are shown in the taxonomy in Figure 3-3 (the physical objects follow the same taxonomic pattern but are not illustrated here) and are numbered to aid the discussion.

3.2.1 Research Data Records (No 1)

Central to the research activity, and perhaps considered of chief importance as potential for re-use and re-purposing, are the objects which constitute data about the object of research (No 1 in the taxonomy). It is from this core data that new knowledge is gained about the object of interest. For clarity the authors use the term 'Research Data Record' (RDR) to refer to any individual item of data of this sort as might be found as a discrete entity and which is either generated as part of the research or pre-exists and is inducted into the current research. Commonly this would be a file such as a spreadsheet, a text document or an audio file, for example. It will be in the form of the RDR that data will customarily be generated, stored, managed and manipulated as required. An RDR contains data that is descriptive of the research object of interest; its rôle is to provide the basis for inferring new knowledge about the subject.

3.2.2 Research Object Data Records (No 2)

In some research the object of enquiry is itself data or, similarly, it exists in the form of a physical or electronic record. So for example, a document of a particular type might constitute an example of a set of documents which themselves are the research object. This type of data record (No 2 in the taxonomy) is referred to as a Research Object Data Record (RODR). A more complex example of the research object of interest might be an entire process represented by a set of associated digital objects, including those of other types (see for example 'experimental apparatus' referred to below).

3.2.3 Contextual Data and Data Records (Nos 3, 4 & 6)

For the purposes of management, however, research data does not consist alone of the data from which inferences are made to produce new knowledge nor (as in some cases) digital objects which are the focus of the research. This key data is often accompanied by contextualizing data of a number of sorts. For example, the direct results of a data collection activity (a research data record) might be illuminated by a description of the methodology, an explanatory narrative or environmental data without which correct interpretation of the key data might be impossible. In addition to this is the sort of information that is common in research and which provides context to research activities, such as standards, papers and other myriad background information on the subject, and information of a similar sort that illuminates the core research data.

Much data provides contextual support as an unintended property and indeed, it is often impossible to tell categorically whether data held in a record has this contextualizing rôle. An example of this can be found in the coding scheme used to classify data and generate the Nvivo document which is part of the research-generated data shown in the Research Activity Information Development (RAID) diagram in Figure 6-5. From the researcher's point of view, the purpose of this data was to provide a means of classifying information drawn from the hand-written notes made during observation of the subject. The outcome of this would be core research data (that is, describing the object of research in some way). However, it is clear that post-project, having this record, together with other data or records with which it is associated and influenced, would provide 'context' and be helpful for the purposes of re-use and re-purposing. This type of unintentionally contextualizing data is shown as No. 3 in the taxonomy, but clearly any instance of this type will also be identifiable as one of the data records having an intentional purpose (No 1, No 2 or No 5 in the taxonomy).

There are some records, however, the rôle of which is expressly intended to 'explain' or illuminate other data which is generated during research. For the subset of research data records which expressly provide context to other data the term Context Data Record (CDR) (No. 4 in the taxonomy) has been adopted.

Some of the intentionally provided contextualizing data may be in the form of metadata which: 'documents the relationships of the content information to its environment. This includes why the content information was created and how it relates to other content information objects' (OAIS Reference Model).

The sort of data that describes the relationship between data or data records is referred to by the authors as 'association data'. Where data associating two or more data records is found in a separate data record such data constitutes metadata, and is referred to as an Associative Data Record (No. 6 in the taxonomy).

Other association data can sometime be found locally in the same file as the data it describes. This might be, for example, data provided in the 'properties' of a text document or consist of a filename which relates one document to another by shared coding. Since such data does not exist as a separate record it is referred to by the authors simply as associative metadata.

Sometimes data or a data record contextualizes an abstract entity, such as the research activity, the subject matter as a whole, or the object of research. Since the entity here being contextualized is abstract there can be no association data making the relationship explicit since there is no concrete object to refer to. Often the content will indicate quite clearly the context; where it does not, for management purposes, it may be necessary to provide metadata that contextualizes this 'contextualizing' data. Such needs reinforce the desirability of using metadata in the management of research data.

The permutations of data and data records which provide context can be seen in Figure 7-1.

3.2.4 Experimental Apparatus Data Record (No 5)

It is also the case that there can exist within the ambit of a research activity, symbolic representations which are analogous to the physical experimental apparatus familiar in much laboratory-based research. This might take the form of computer code and controlling parametric information (analogous to a lab rig and its physical parameters) which processes input data and generates output data. This is identified as No. 5 in the taxonomy and referred to as Experimental Apparatus Data Record (EADR). Figure 3-4 compares an example of a conventional lab-based experiment with one in which all the elements of the research are 'data' of one sort or another.

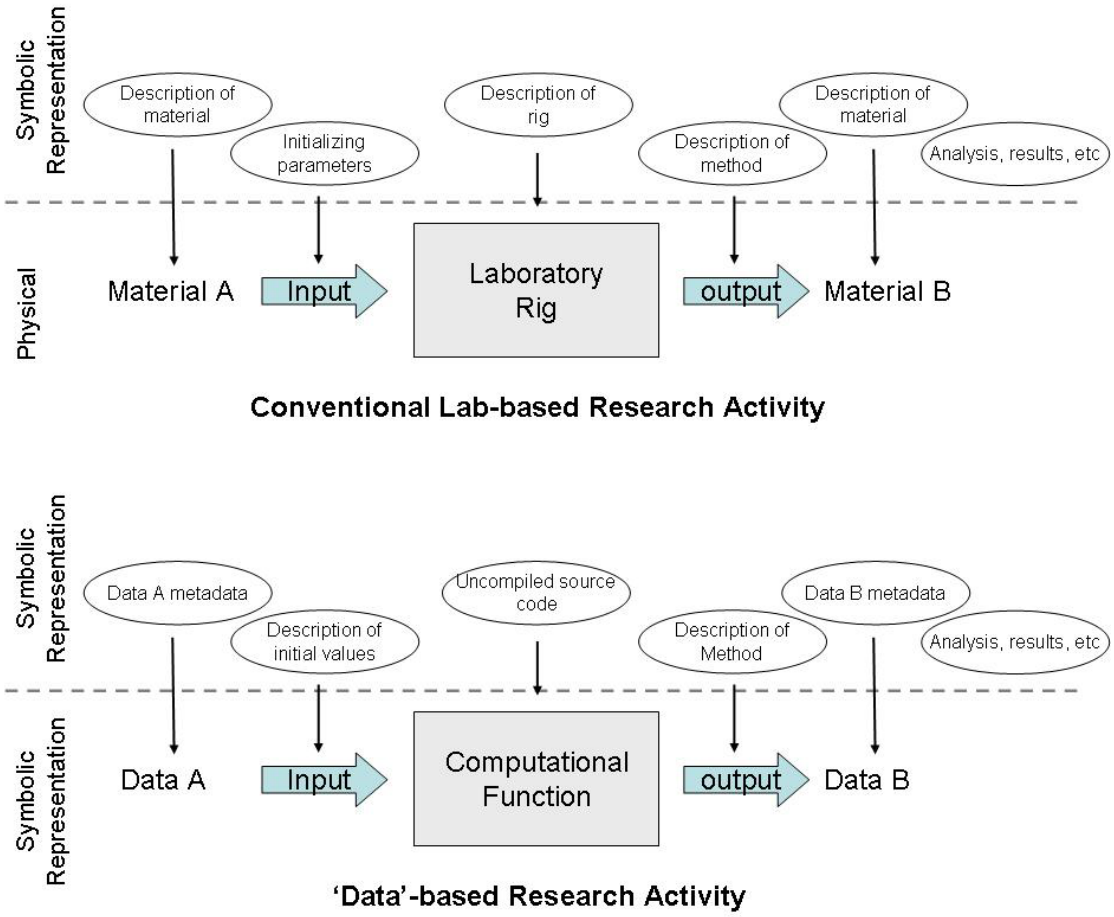

Figure 3-4. A comparison between an example of a research activity carried out in the physical domain and one solely carried out in symbolic representations

It can be seen from the above that it is possible for an entire research activity to be constituted and recorded in symbolic representations, in which all six of the different types of data object can be found. However, it has been found by the authors when in the analysis of the data case audits reported in Section 6 that there is often real difficulty in identifying what rôles particular instances of research data have played in the activity. This problem is discussed in greater detail in Appendix C.

3.3 Levels of Management

So, data themselves can be managed (that is, within the record) as can records. In addition to this it is possible to see management at the 'case' level. For example, in a single experiment (perhaps those shown in Figure 3-4) there will be a number of research data records, contextual data records and others which are logically associated and which 'tell the whole story' of the one experiment. The absence of one or more of these files, or loss of the information that these files are related, may make interpretation of the remaining information impossible. Such collections of case will, too, be the subject of management, and will be of greatest value only whilst their association at the 'super'-case level is retained where not only are intra-case associations recorded but so to are inter-case associations.

3.4 Identifying Objects

The provision of the taxonomy of data objects might suggest that identifying which rôle a particular data record fulfils will be self-evident. Scrutiny of existing research cases suggests strongly, however, that it may be very difficult to identify in any particular activity or project which records are inputs and outputs to the research, which might be directly analogous to experimental apparatus, which controlling parameters, which core research-generated data and so on.

There is some evidence (based on the research data case audits) to suggest that there is a hierarchy of difficulty in identifying what rôles different data records fulfil which is dependent on the research activity type. Blessing & Chakrabarti (2009) introduce a design research methodology which distinguishes between *descriptive research* on the one hand and *prescriptive research* on the other. Scrutiny of the research data cases audits suggests that post-project identification of the rôles of data records is quite straightforward in descriptive research projects, and more difficult in prescriptive research projects. In particular, identification is especially difficult in prescriptive studies where the research object(s) (in the form of RODRs) are research generated. An example of can be seen in the RAID in Figure 6-2 . Here it is difficult to know what constitutes an RDR, a CDR and RODR, etc. In one sense, all the records appear to act as contextualizing data for the other records in the case and some could be said to stand together as a proxy for the research object of interest.

It may be possible as a result of further analysis to provide a management methodology in which the approach to management is based on the type of research that is being carried out.

However, an alternative, precautionary, approach could be to advise that all related data records in a research activity (i.e. a research data case) be subjected to management indifferent to the rôles that they played. At the same time, given the special understanding a researcher would have of his own data, it would seem wise to ensure that as much management as possible be done for re-use and re-purposing during the duration of the research activity. Having said this, as reported below in Section 4.3.1, researchers themselves sometimes had difficulty in identifying and classifying their own data.

3.5 Modelling the Data Objects and their Relationships

A number of relationships are suggested and can readily be observed between the objects and the levels at which they are associated. The levels of management and the relationships are shown graphically in Figures 3-5 to 3-7 below. The figures constitute the basis by which the actual development of data during the research activity can be mapped, as discussed in Section 5.4. For simplicity's sake the objects shown in the figures are limited to RDRs and CDRs; in principle however, any of the six main category of data records may feature in a relationship diagram.

In Figure 3-5, what the authors refer to as a *horizontal* relationship can be observed. This sort of relationship occurs as a result of specific and appropriate processes being carried out on data at the data level, resulting in data change or generation and often the gestation of new records. Characteristically, these processes include such things as data *refinement*, data *derivation* and data *aggregation.* Sometimes these activities result in

new RDRs, sometimes in changes which occur within an existing RDR (for example, the addition of a chart or a new page to an existing spreadsheet).

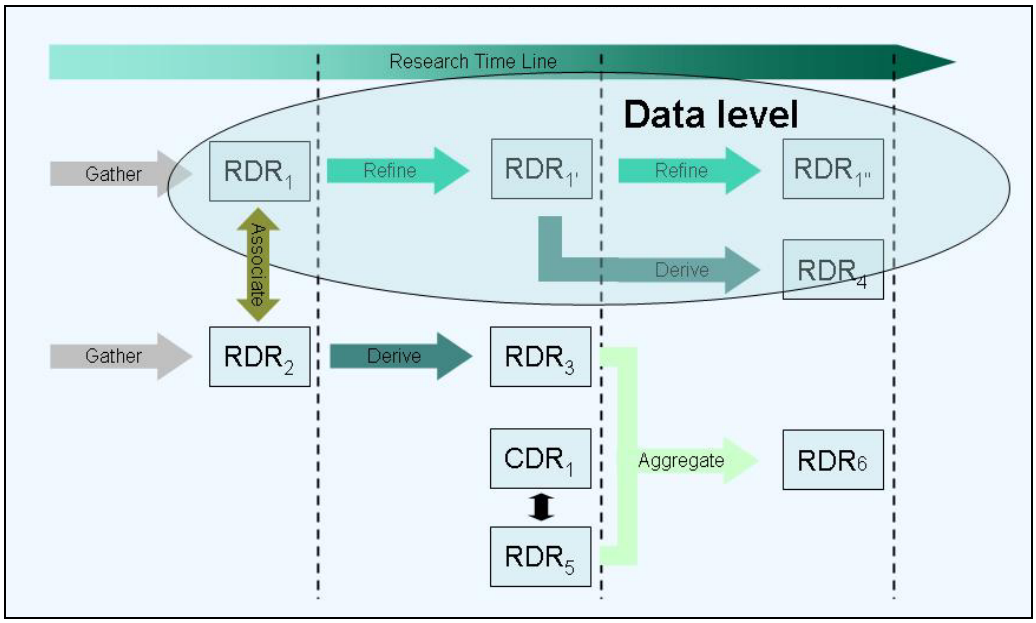

Figure 3-5. An illustration of the *horizontal* relationship between data records occurring as a result of processes being carried out at the data level

Additional to this is the sort of relationship which occurs between data records. This the authors refer to as a *vertical* relationship (see Figure 3-6). This sort of relationship is either inherent or developed as a result of specific and appropriate management sub-activities being carried out at the data record level.

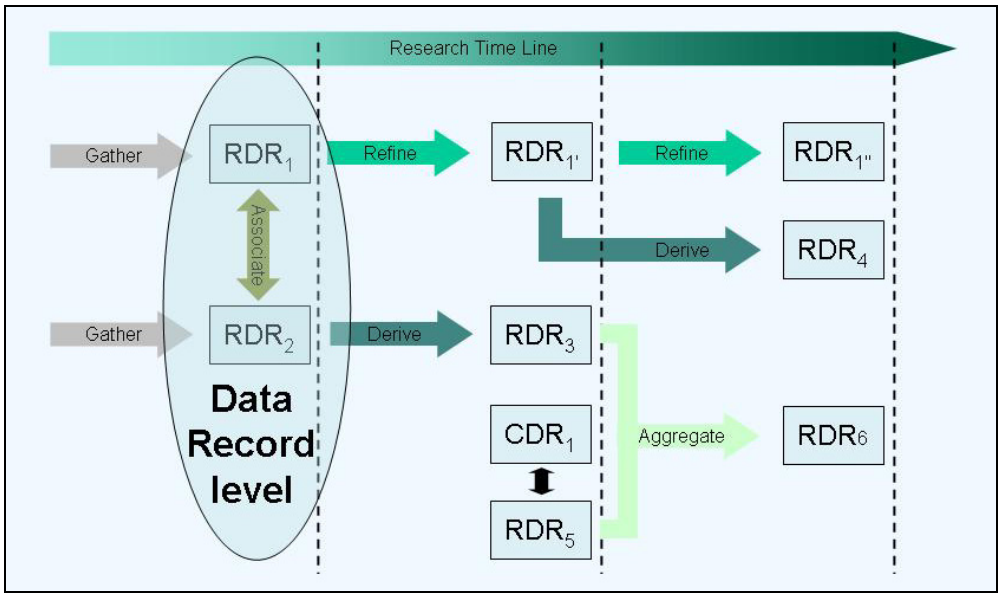

Figure 3-6. An illustration of the vertical relationship that can exist at the data record level.

Characteristically these activities include such things as *generation, collection, association.* These relationships can exist between not only RDRs but also between CDRs and between CDRs and RDRs. Where these vertical relations are implicit only, good management practice would require explicit *association* by some mechanism. For example, in a research project it is often the case that a number of similar experiments will be made, each one of which will have a 'case' of records in which research data is contained (see Figure 3-7). Such collections of cases will, too, be the subject of management, and will be of greatest value only whilst their association at the 'super'-case level is first stated, recorded and perpetuated.

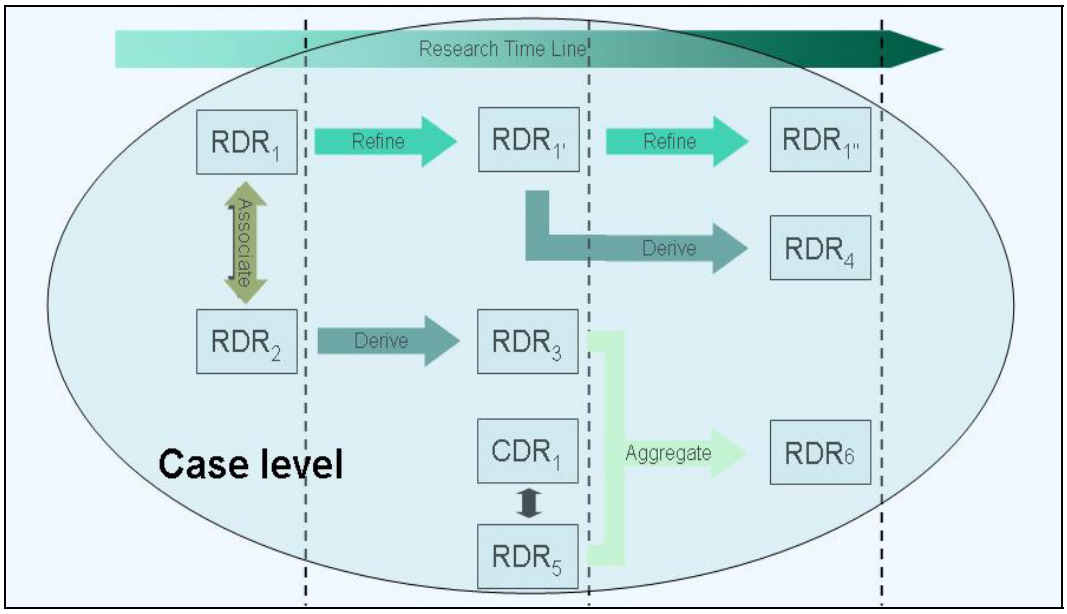

Figure 3-7. An illustration of the vertical relationship that can exist at the case level

The above considerations provide an understanding of the 'terrain being mapped' and the basis for discussing it. Given this, it is now possible to consider the ways in which the data preparation activities and the data development processes can have a bearing on management as discussed in the following section.

3.6 The Implications of Data Preparation and Development for Management

The activities involved in managing research data are those identified as being associated with the sub-activities of *curation, archiving* and *preservation*, as defined by Lord and Macdonald (2003) and as discussed in Ball (2010). Curation, a relatively newly coined term in relation to both data and digital artefacts, is defined as:

> "The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials."

Amongst others the tasks identified in the above works, and generally, as being appropriate to curation are: selection of the data sets to curate; bit-level preservation of

the data; creation, collection and bit-level (or hard-copy) preservation of metadata to support current and future use of the data; the storage of the data and metadata with levels of security and accessibility appropriate to the content; the updating of data sets; and the transformation of the data to allow compatibility with previously unsupported and new work flows and processes.

Whilst the motivations of the curator of data and the motivations of the researcher may well be different, these curation-related activities are of exactly the sort that the researcher and research manager will be engaged in during research if data is to be made available for current and later use. At a more microscopic level, management includes such things as organization (disposition and classification), collation, association and so on, in addition to the other activities which have been identified above as more closely related to data development in relation to the research – rather than the curation – process. Some or all of these activities can occur, according to the logic of their definitions, at the data level, the data record level and the case level. The incidence of processes in relation to each level is suggested below in Table 3-2.

Table 3-2. The levels at which data development occurs.

Development Processes	Data Level	Data Record Level	Case Level
Addition	✓	*	*
Association	✓	✓	✓
Aggregation	✓	*	*
Annotation	✓	✓	×
Augmentation	n/a	✓	n/a
Collection	✓	×	×
Collation	×	✓	✓
Deletion	✓	✓	✓
Derivation	✓	×	×
Duplication	✓	✓	✓
Extraction	✓	✓	×
Generation	✓	✓	✓
Migration	✓	×	×
Population	n/a	✓	n/a
Refinement	✓	×	×

Note that those entries coded by ‘n/a’ are cases where an occurrence is precluded by virtue of the definition of the term and therefore not applicable, those coded by an × are cases where an occurrence is logically precluded.

For example, it makes sense to say that data can be associated with other data, as can data records with other data records. It does not make sense, however, in light of the

definitions, to say that a research data case can be derived from a research data case or come about as a refinement.

Particularly at the data level, these activities may produce one or more of a number of side-effects, requiring some sort of management intervention if the aims of curation are to be supported and achieved. The side-effects are not only a logical contingent of data development activities, but also can be demonstrated in specific instances of development.

3.6.1 Management Side-effects

Consideration of the development processes and scrutiny of the data audit cases (treated in Section 6) suggest to the authors that five chief side-effects occur as a result of development, these being:

- information loss
- information gain
- function loss
- function gain
- state loss

One or more of these will be characteristic of each of the data development activities. Some examples of these are given below by way of illustration of the side-effects together with likely necessary management interventions:

- **Data Refinement:** The refining of data is commonly accompanied by *information loss*. For example, rounding a real number up or down to an integer will mean that the original level of detail is lost. Clearly, if information is lost in this way from a data set being prepared for a particular purpose, the opportunities for data re-use for another purpose may be reduced. To overcome this it may be necessary to record explicitly the detail of the refinement process or the existence and association of the earlier version of the data. The recording of the process may be particularly valuable if it is *reversible*. If such is the case, retention of the earlier version of the data may be unnecessary; however, sufficient contextual data will be required in order to regenerate the earlier data.
- **Data Derivation.** Deriving new data from that which exists is characterized, paradoxically, by both *information gain* and *information loss*. Information gain is not in itself a problem, after all the purpose of generating derived data is expressly to increase information. Nevertheless, it could be helpful for data re-use if contextual information is available to show where, or in what manner, the gain has occurred. Information loss is clearly more problematic; for greatest re-usability the association between derived data or data records and the raw or precursor data upon which they were based requires preserving. Derivation is the process of creating a new description of the subject described by the existing data (perhaps existing in more than one location) examples being such things as a histogram, a textual commentary, a statistical profile, a narrative report and so on. At the very least derivation means a loss of detail, often it constitutes a complete loss of the original data or the loss of association between data.
- **Data Migration.** In migrating data the express intention is to transfer digital information from one format to another, yet with the intention of preserving the full

information content. (This may be required for the purposes of curation, perhaps mitigating the effects of technology obsolescence, or for portability during use.) Nevertheless, this development process is characterized both by *information loss* and *function loss*. As observed by Beagrie & Jones (2001) in migrating it is not always possible to make an exact digital copy or replicate original features and appearance and still maintain the compatibility of the resource with a new generation of technology; likewise with a different technology. The migration of data from a live (i.e. computationally supported) spreadsheet to a PDF format provides an illustration where, for example, metadata may be lost in the transformation (information loss) as will the underlying spreadsheet functionality (function loss). These are quite gross losses; often the losses will be more subtle, yet good management requires that the losses be understood and mitigated.

- **Data Deletion.** The deleting of data is a perfectly normal activity in research data development and management. Nevertheless, erasing or obliterating data may have its dangers since, once gone, it may be impossible to retrieve or reformulate. An important side-effect of deletion is where data which represent the state or condition of an artefact, subject or system is lost (*state loss*). This may not be important for the current research but may be for potential re-use or repurposing. State loss occurs either as a result of discarding unregarded data or overwriting an existing version or the data. It characteristically occurs when the item of interest is the final outcome of an iterative process – for example in the application of the Delphi Method – or when continuous updating of data occurs, for example in the automatic updating of computer code in response to closed-loop feedback. Consideration of future use of research data may modify the researchers' approach to the discarding of data.

As noted above these side-effects suggest the notions of process reversibility and repeatability. Clearly, where a process is carried out on data that transforms it in some way, it may be necessary either to associate the new data with the precursor data (where the process is not reversible) or provide the means by which the earlier data can be regenerated or recovered. In the second case it would be necessary to provide the reversing function and parameters as part of the contextual metadata. An illustration of non-repeatability might be where a subjective coding scheme (i.e. where a classification judgement has to be made by the researcher) is applied to raw data to create a derived set of data; because of its subjective nature this may not be exactly repeatable.

3.7 Overview

In this section theoretical consideration has been given to the processes and activities associated with the research activity. A method of modelling the development of research data has been derived which will allow the development of data within a particular research project or activity to be mapped. At the same time the side-effects of such development and their implications for data management has been revealed.

In addition, these theoretical considerations and discussion provide a framework by which an enquiry of research data in the engineering domain and its better management can be structured. The remaining sections of this paper discuss work in which research data representative of that found in engineering research is selected, audited and characterized. This empirical work serves to support much of the theorizing above.

4. THE SCOPING SURVEY

The purpose of the scoping survey is to identify a spectrum of engineering research information and data. From this broad spectrum, distinctly different engineering research data records are to be identified. As an additional aim, each of the data records is to be classified by a set of attributes. This scoping work will in due course be used as the basis for selection for audit a set of data which is representative of and characterizes the spectrum.

4.1 Sources of Data Records

The authors selected as a source of data for their study two repositories. The first of these repositories is constituted of the data assets held by researchers in the Innovative Design & Manufacturing Research Centre (I*d*MRC) located in the University of Bath. The data available is that associated with the I*d*MRC's research projects over a period of nearly a decade, a representative sample of which had already been subject to an audit using the DAF methodology (Jones, et al, 2008).

The second repository is 'virtual' in that it consists of an inventory of data assets associated with a large-scale research project distributed geographically over eleven universities. The project in question (the KIM Project) was of a highly inter-disciplinary nature covering a range of engineering research topics looking at such diverse things as product modelling, document management, information archiving, information modelling and engineering standards, the nature of learning organisations and HR policy for product-service systems. The diversity of the subject matter is reflected in the data assets.

First, the researchers identified candidate data assets using certain practical criteria such as whether the data asset is still available or whether the originator or owner is still amenable. From these candidate data assets, 12 cases were selected which the authors believed display a good mix of research topics, methods, data size and data format.

4.2 Selected Data Assets

The assets for the scoping survey are listed below, consisting of 7 assets from the I*d*MRC data audit (Section 4.2.1) and 5 from the KIM inventory (Section 4.2.2).

4.2.1 From IdMRC Data Assets

The following data assets were selected from the IdMRC inventory:

1. Airframe Stress Data Reuse
Industrial material description tool
Process flow (Visio)

2. Snow Mobile Design Activity Observation
Specification documents
Information resources
Analysis Information
Output:
- Video clips
- Design activity ouputs

Activity record
Product model
Design Records:
CAD models
XSLT
Topic maps

3. Aerospace Cost Forecasting
Bill of materials
Manufacturing dataset
Survey on cost estimation practices
Serenity Vanguard Library

4. Large-Scale Metrology Shared Resources
Standards library
Indoor GPS/Laser tracking measurement benchmarking data
Laser Scanning point cloud data
Laser scanner measurement evaluation data sets

5. Form-fill-feed Packaging Modelling
Variations in product size
Product Flight Path

6. CNC Machine Measurement
Machining and measurement process plan file

7. Cryogenic Machining
Shoe sole data:
Models of 3 solutions
Student projects
Soles to fit patient prescription
Machining parameter data:
Thermal data profiles
DMTA data
SEM data
Surface roughness data
Machining data
Force calculation data

4.2.2 From the KIM Project data assets:

8. Information Management Tool
Requirements Interviews
Audits + supporting /contextualising word doc; Output some facet docs a facet workshop and concept maps

9. Knowledge Enhanced Notes
Data source + meta data linkage

10. Service Design Research
Power Delivery case study 1 data:

Story lines
Survey on risk at the different stages of the proprietary process
Question template
Responses
Captured data in a excel file (manual data entry into excel)
Key service data + supporting evidence

Power Delivery case study 2 data:
Case study data
Requirements to establish reliability of supplier

11. Design Activity & Knowledge Capture Research

COSTAR:
Log files
Video capture of the activity
$IDEF_0$ maps
Story board which is a combination of them.

Bamzooki:
Log files

12. Understanding the Learning Organization

Interview guides
Voice recordings
Interview transcripts
Field note observations
Meeting Transcripts
Documents from subject company

Aerospace Contractor:
26 interviews and observations (interview transcripts, field notes)

Defence Contractor:
22 interviews and observations (interview transcripts, field notes)

Power Conversion Contractor 1:
31 interviews and observations (interview and meeting transcripts, field notes)

Power Conversion Contractor 2:
23 interviews and observations (interview and meeting transcripts, field notes)

4.3 Scoping Survey Inspection

For each of the above selected assets a semi-structured interview was carried out with the principle researcher associated with the asset. This interview had two stages, first to identify individual data records (Section 4.3.1) then assigning attributes to each of the data records (Section 4.3.2).

4.3.1 Identifying data records

Identifying the different data records was a much more difficult task than anticipated. In many cases it was hard for the researchers being interviewed to identify what they considered to be their 'raw' data (that is, data that has not been subjected to any development during the research activity). There was also difficulty in some cases defining the boundaries of data records. Partly for consistency, a data record was considered to be a single electronic file or multiple files all of the same type with the same purpose (e.g. a stack of questionnaire responses was considered as one record).

The second issue concerns what it is that distinguishes a *research* data record from a non-research data record. Unfortunately, this was not a binary choice as there was more a continuum of data records that had a varying relation to the research activity (see above for the theoretical viewpoint on this). After much consideration it was decided that the records not directly related to the research activity would not be investigated any further. It seemed that of the related data records two classes could be formed; those records that contain data that was the object of inquiry (termed Research Data Records) and those which provide context by which to understand or interpret the research data records (termed Context Data Records). There would not necessarily be found an explicit association between these types of record. (Note that this dichotomy of record type would be elaborated as the research developed into a classification of six record types which more fully describes the variation in records to be found in research – see Section 3.2 for details.)

There is another area of uncertainty, which arose when subjecting the project 'data' to scrutiny and classification (both in the scoping study and later in the data case audits), which requires further consideration and arises in those research activities in which the object of research scrutiny or the tools used for exploration are themselves data-like. A difficulty arises in being clear about what constitutes 'research data' and what constitutes material analogous to 'experimental apparatus'. This difficulty is discussed in some detail in Appendix C.

In total 46 data records were identified and classified.

4.3.2 Assigning attributes to data records

Once a data record was identified, a series of questions were asked of the interviewee in order to assign attributes to it. The questions were arrived at by inspecting various documents and data sets. Questions were suggested by this process and those that seemed to be both interesting in terms of management ramifications and characterising data to distinguish between data sets were used during the interviews.

The interview method, consisted of the following questioning strategy:

1. Ask the question and allow for the participant's own interpreted response.

2. Give examples of the type of responses we were looking for and ask them to re-evaluate in the light of this extra knowledge.

3. Give the answer we anticipate and discuss.

The questions asked in order to classify each data record are given in Appendix B.

The responses to the questions were recorded in a database in the manner shown in Figure 4-1.

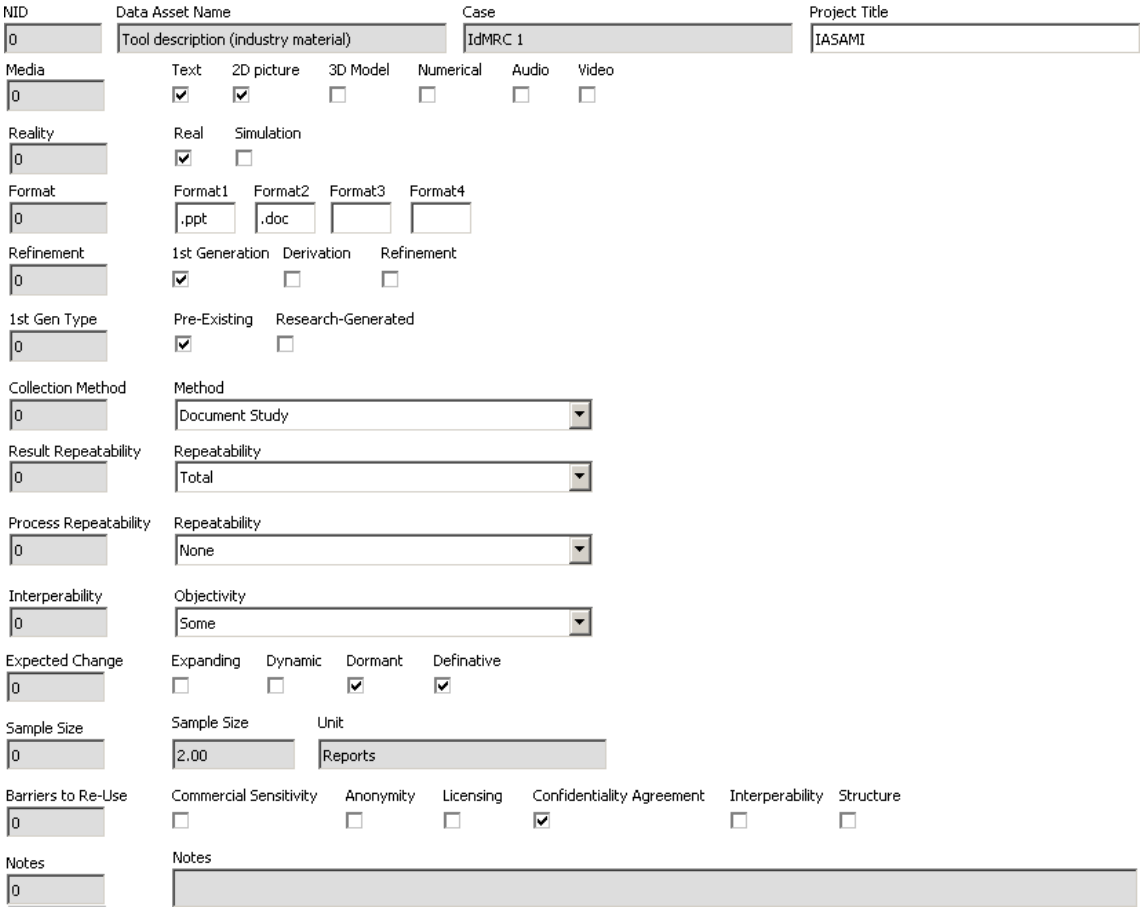

Figure 4-1. Example Scoping Survey Response Record

4.4 Findings from Scoping Survey

The scoping survey revealed that some of the questions asked were far more revealing in terms of characterising the data records than were others. It was also found that not all of the questions had meaning in all cases or to all participants, particularly with regards to the question on repeatability. It seemed that certain data records were associated with research performed by means of prescriptive study, which is quite common in engineering research. For these data records many of the questions had less pertinence than when they were asked in respect of data records of descriptive studies. The authors hypothesize that this is a key distinction, one which was to be included in the case study selection criteria.

5. DATA CASE AUDIT METHODOLOGY

The following section will discuss the methods chosen for the case study inspection and audit, starting with discussion of the level of granularity at which the inspection would be conducted.

5.1 Selecting Case Studies Candidates from the Scoping Survey

Having entered the scoping survey result into a relational database it was then possible to run queries based on selected attributes of the data records in order to select the case

studies. It was reasonably clear that completely distinct groups were not formed as many had overlapping attributes.

The aim was therefore to identify outlying cases, which would allow a selection of data which displayed the greatest variation across the attributes which had been used in the original data characterisation. This was done by partitioning the data assets by using the following classifications from the scoping study characterisation: generation type, media, study type and reality, as shown in Figure 5-1 as a series of four distinguishing steps. In Step 1 the data records were separated into those which were generated during the research process and those which were collected from an industry source, this being *pre-existing data*. In Step 2, all the data that was *research generated* was further classified into that which was of the same media type and that which consisted of highly variable media type. In Step 3 all the records which were *highly homogenous in media type* were further separated into two classes, one of which represented descriptive research and the other prescriptive research. The final step, Step 4 classified the data records from *descriptive research* into 'real' data (sourced from real world activity) and 'simulated' data (resulting from a simulation). Resulting from this partitioning five separate groups were identified and from each group, one case was selected for full case study audit thereby maximizing diversity across these attributes. The heuristic adopted for the case selection was to pick from those available the case that had the greatest explicit internal diversity of data types. The selected cases were: aerospace cost forecasting; snow mobile design; design activity and knowledge capture; understanding the learning organization; and large-scale metrology (see Section 4.2 for more details of each case). The authors acknowledge that there are other ways of partitioning data in such a way to maximize diversity, each of which would arrive at a different selection. However, it is difficult to say how these different approaches might be compared and benchmarked.

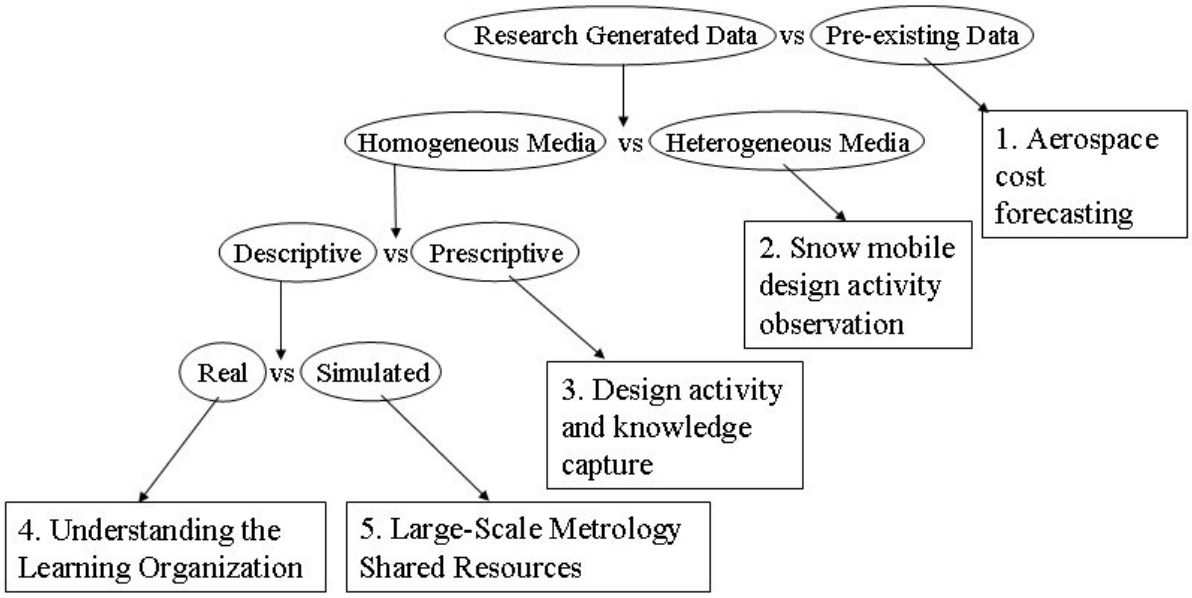

Figure 5-1. The selection of data cases for audit

Table 5-1 lists and numbers the inspected case studies and also labels the section and their associated sections of the report.

Table 5-1. List of data case studies for audit

Case #	Description	Section
1	Aerospace cost forecasting	6.1
2	Snow mobile activity observation	6.2
3	Design activity and knowledge capture	6.3
4	Understanding learning organisations	6.4
5	Large-scale metrology	6.5
6	Cryogenic machining (trial case)	6.6

5.2 Granularity of Inspection

It has been observed earlier in this report that much of the research conducted into research information management has been at a coarse level of granularity. Such research predominantly concerns high-level information management activities such as backing-up, file structures, file-naming conventions etc. It was identified that there was a gap in research for supporting information management which occurs at the 'create or receive' level of the DCC Curation Lifecycle Model (Figure 2-1). It was therefore decided that the best level of granularity at which to inspect each case study was to map for each case what the authors now refer to as the Research Activity Information Development (RAID), based upon the model of the relationships and the development processes discussed and illustrated in Section 3. This would allow capture of research data and information at all three levels (case, record and data) as described in Section 3.2 whilst identifying the implications for management of data preparation, developments and side effects (discussed in Section 3.6).

5.3 Method of Capture

Meetings were arranged with the principal researcher in each of the case study candidate projects. The meetings were recorded using a webcam and a video camera. The candidates were asked to assist in the creation of a RAID diagram using a white board and sticky notes (see Figure 5-2 for an example). The sticky notes were used to describe data records involved in the particular case. The arrows between sticky notes were used to describe data preparation and management activities as described in Section 3.6. The process was quite organic but tended to start from a sticky note describing what was considered to be the main raw data set of the case study. The diagram was then expanded by the data owner with some guidance and questioning from the researchers involved in the ERIM Project. The terms used to describe each of the data records and the data preparation and management activities were captured in the terms used by the data owner. The diagram was then formalised using a standard modelling representation language in conjunction with the Data Management Terminology introduced in Section 3.

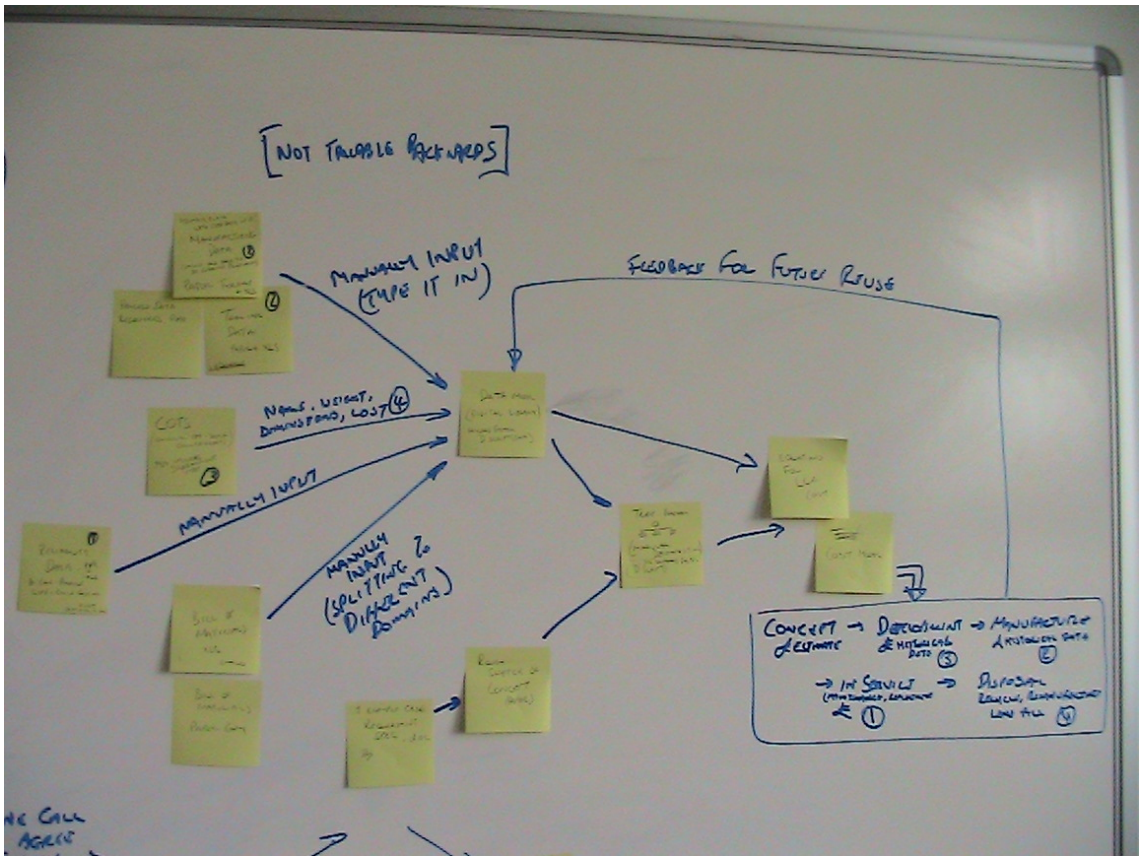

Figure 5-2. Example RAID at the point of construction

5.4 Modelling Research Activity Information Development

Only one other example of research has been found that attempts to model research information flow, at a similar granularity, this being the research conducted by the Research Information Network (RIN) and the British Library (Williams & Pryor, 2009). That research modelled patterns of information use and exchange in several life-science case studies. A concern about that work is a lack of consistency in the representations, such that difficulty occurs in comparing cases.

It was therefore decided that a different approach should be taken, where consistency could more easily be achieved by using the modelling approach introduced in Section 3.5 – which combines relationship and development information as well as distinguishing between data record types – used in combination with a formal standard representation language.

A number of such languages were reviewed of which both IDEF3 and UML were found to be suitable, each having their own strengths and weaknesses. Of these two, UML was chosen principally because it has been used in the area of research information management before by the OAIS Reference Model (CCS01, 2002). It was thought that adopting a formalism for such work already familiar to those in the community would be helpful.

6. CASES OF RESEARCH ACTIVITY INFORMATION DEVELOPMENT

The following section describes a number of research cases under inspection and shows examples of the first attempt at modelling the information development for each case based on the underlying development activities. It should be said that these attempts at modelling Research Activity Information Development (RAID) represent an exploration of the technique, which is being revised and matured as the research continues. However, notwithstanding the immaturity of the modelling approach, it can readily be seen that the method allows representation of the way that data is developed and the relationship between records in a way that has not been possible before.

In addition it is possible to denote within each RAID diagram data records associated with either 'descriptive' or 'prescriptive' elements of a research activity. To indicate this the groups of research data records associated with each mode of a research activity are enclosed within boxes, labelled either 'description' or 'prescription' as appropriate. The distinction between and importance of understanding the differences between descriptive and prescriptive research is treated in detail by Blessing & Chakrabarti (2009). Descriptive data is said to describe or constitute the research subject, whilst prescriptive data describes or constitutes the proposed change or improvement proposed or implemented given the research findings. 'Description 2' is constituted of research records of work which validates the 'improvement' recommended in the prescriptive study. In 'D2' studies, the prescription is considered in the new situation and is evaluated (an example can be seen in Figure 6-3).

The case studies are presented here by way of illustrating the modelling approach rather than, at this stage, as a detailed analysis of the management implications. This aspect will be the focus of later reporting.

6.1 Case Study 1: Aerospace cost forecasting

The following case study was chosen as a good example of where researchers may take pre-existing data to use for their research. In modelling this study two quite distinct RAID diagrams emerged (see upper and lower boxes in Figure 6-1), linked only by the fact they were associated with a similar topic of research conducted by the same researcher. This separation suggests there to be no dependence or precedence between data represented in the upper and lower elements of the diagram. The lower case describes a simple survey and was not of particular interest other than as a record of the survey's existence. The upper case shows a more complicated RAID diagram consisting of similar data sets from four separate companies/projects.

What is of particular interest is that the actual data in the records (represented within the 'industrial project data' dashed box) – notwithstanding the fact that trouble had been taken in their collection – have no importance to the researcher. What is being researched is whether the cost modeller that processes this data works as intended.

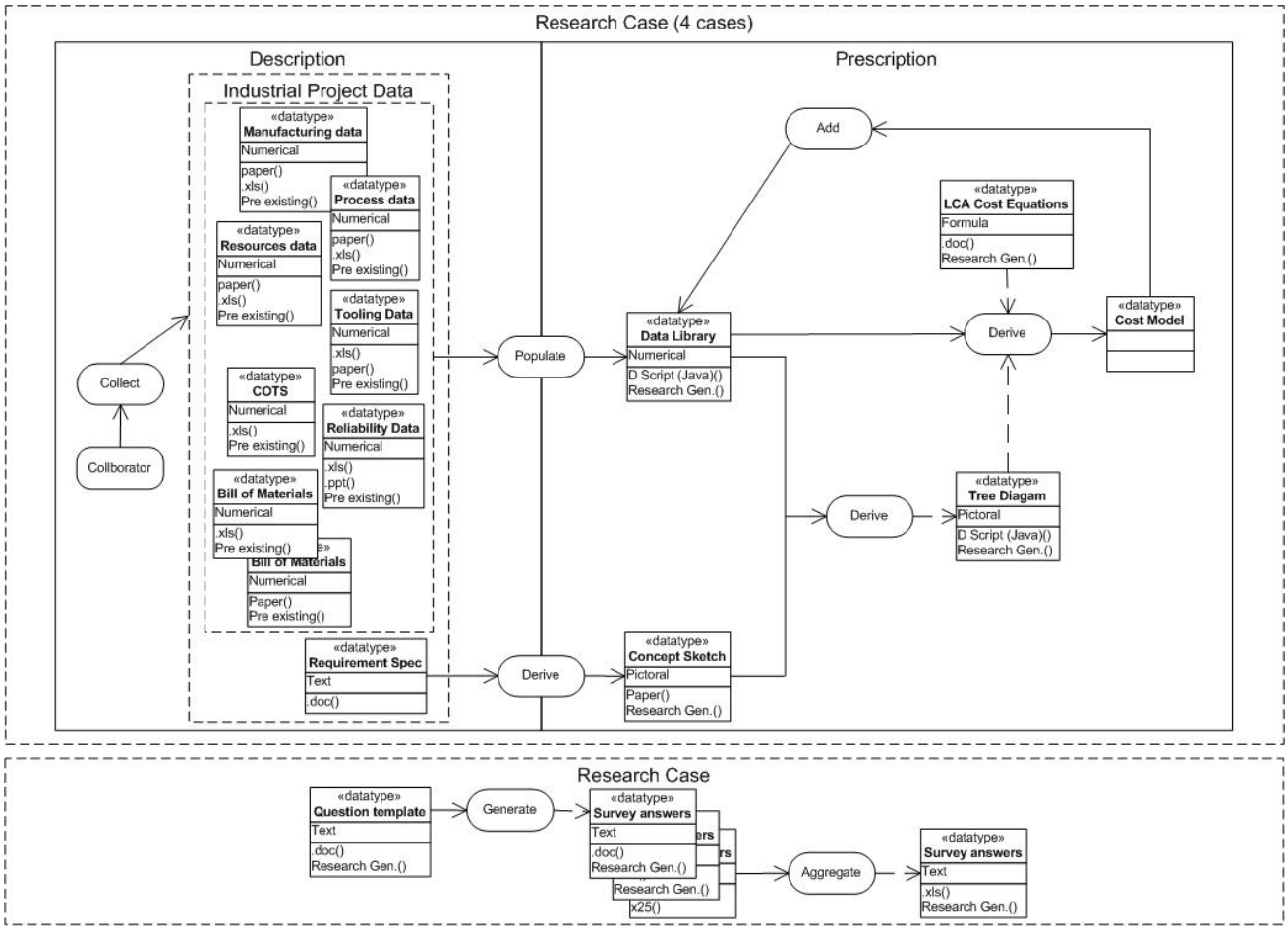

Figure 6-1. Aerospace cost forecasting RAID

6.2 Case Study 2: Snowmobile design activity observation

The following RAID describes a prescriptive study. What makes this RAID diagram particularly complex is the existence of multiple instances of data record 'generation', which make the research data harder to interpret.

Also of interest is the fact that two of the first-generation data asset sets (identified as analysis files and selection/decision data) seem to play a part in both a descriptive study and then a prescriptive study, and that their rôles in each (and therefore their relationship with other records and the 'context' they provide) would be different. In the descriptive study these records would constitute *generated* data (an output) and in the prescriptive study *raw data* input to a 'new improved' process. This research involved creating a topic map representation from a stage design activity in an attempt to provide a more satisfactory means of recording the design process and its rationale.

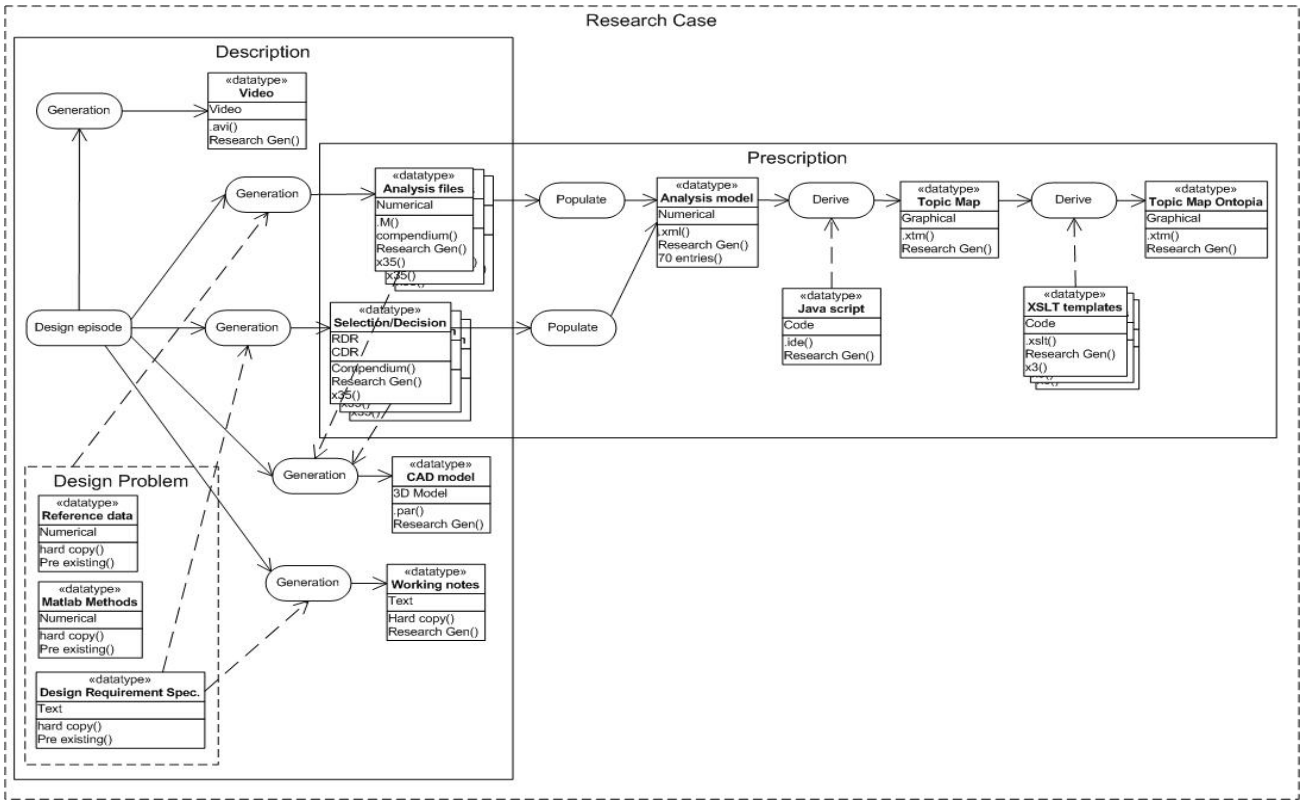

Figure 6-2. Snowmobile design activity observation RAID

6.3 Case Study 3: Design activity and knowledge capture

This case is very similar to the previous one. Here design in a virtual environment is captured and represented in two different formats (IDEF0 and DREd) as a prescriptive study. Inevitably some descriptive material is generated during this trial implementation of the prescription and some descriptive material is created to evaluate the different forms of representation by means of a questionnaire.

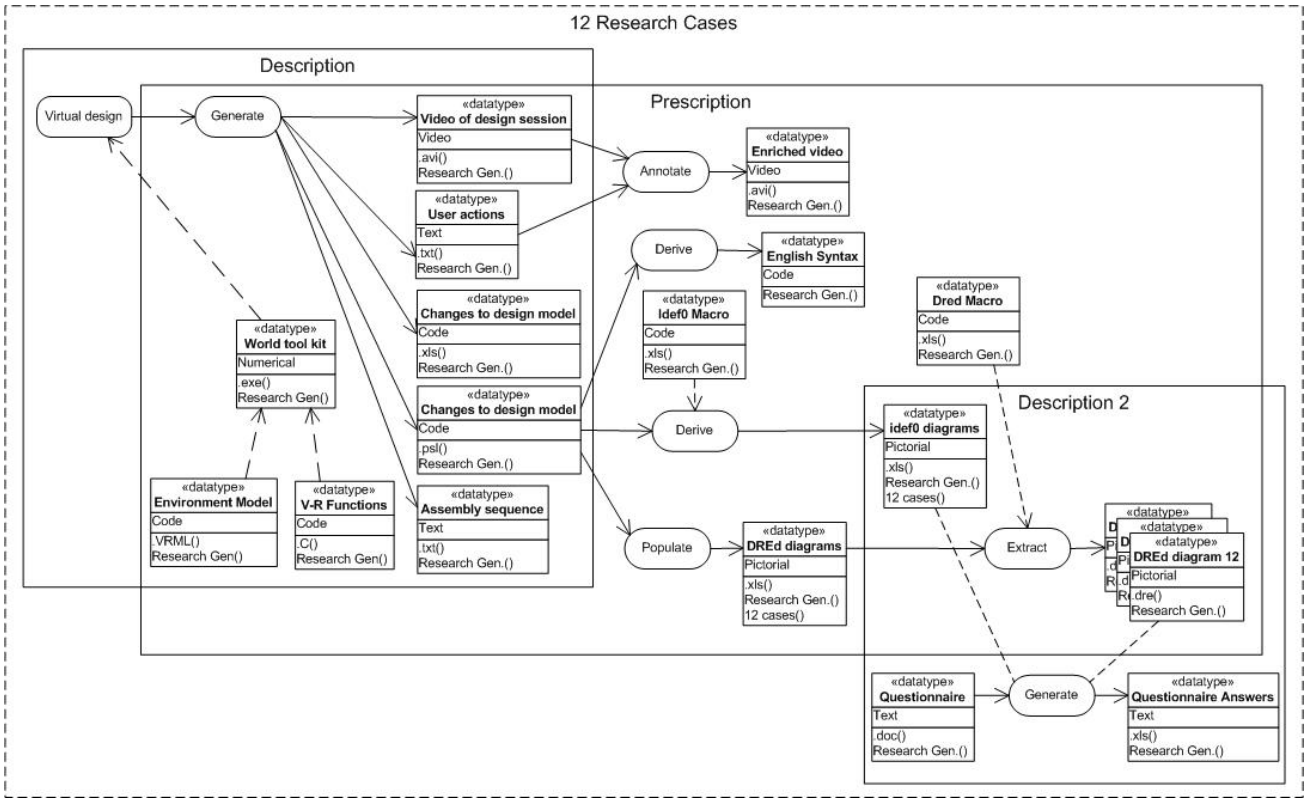

Figure 6-3. Design activity and knowledge capture RAID

6.4 Case Study 4: Large-scale metrology

This study shows two distinct RAID diagrams. The upper diagram details a new 'prescription' to report the uncertainty of a measurement system by varying the positioning of the measurement equipment. The lower shows a new method of visualizing, for the purposes of back-engineering, the data found in the input files.

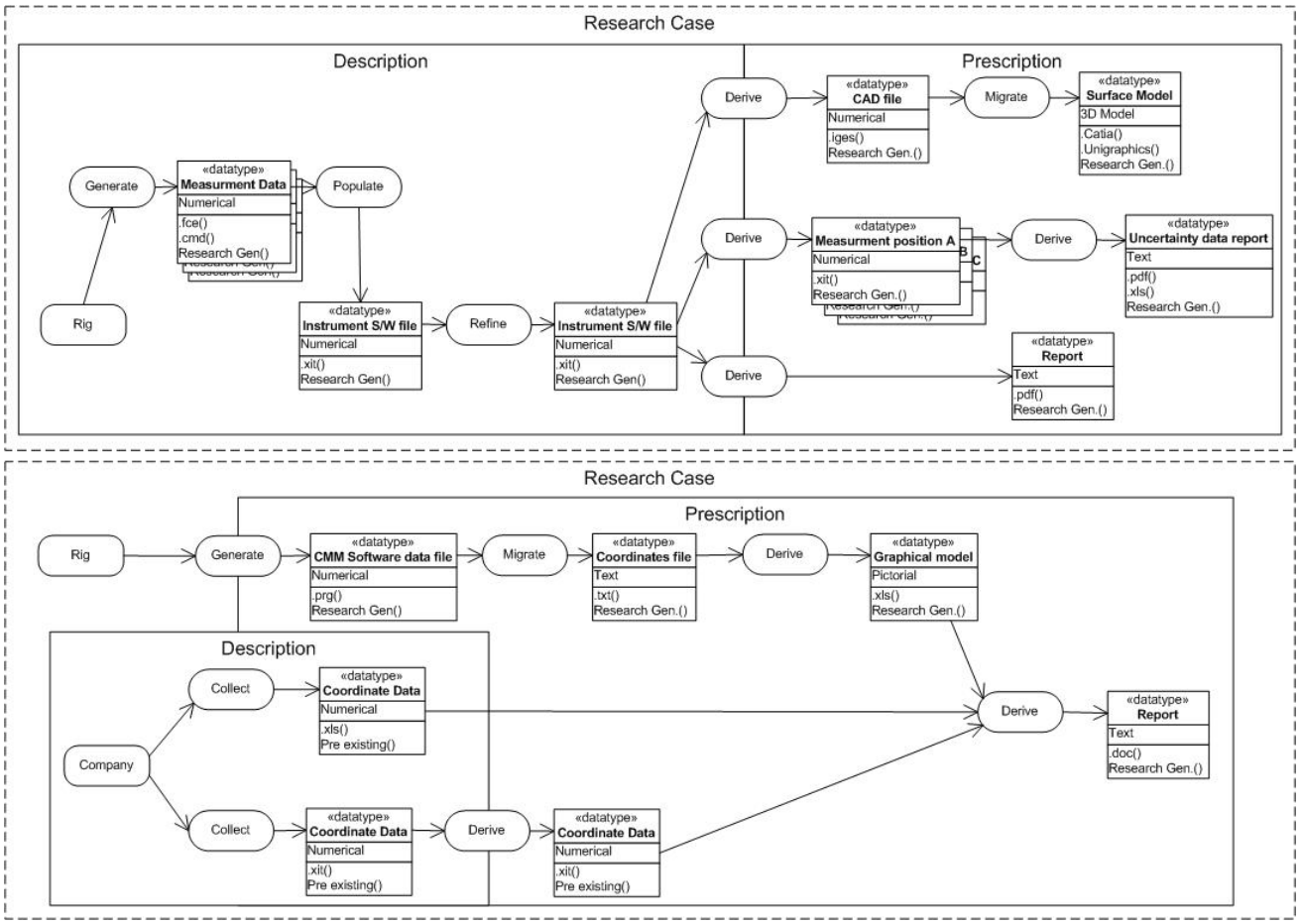

Figure 6-4. Large scale metrology RAID

6.5 Case Study 5: Understanding the learning organisation

This is a purely descriptive study taking similar data from interviews, observations and meeting and coding them into an Nvivo document and field notes. Though it was thought potentially easy to trace back from the Nvivo coding to the first-generation data, the reports found at the end of the study had few links to the raw data, indicating that the study involved a large amount of holistic interpretation.

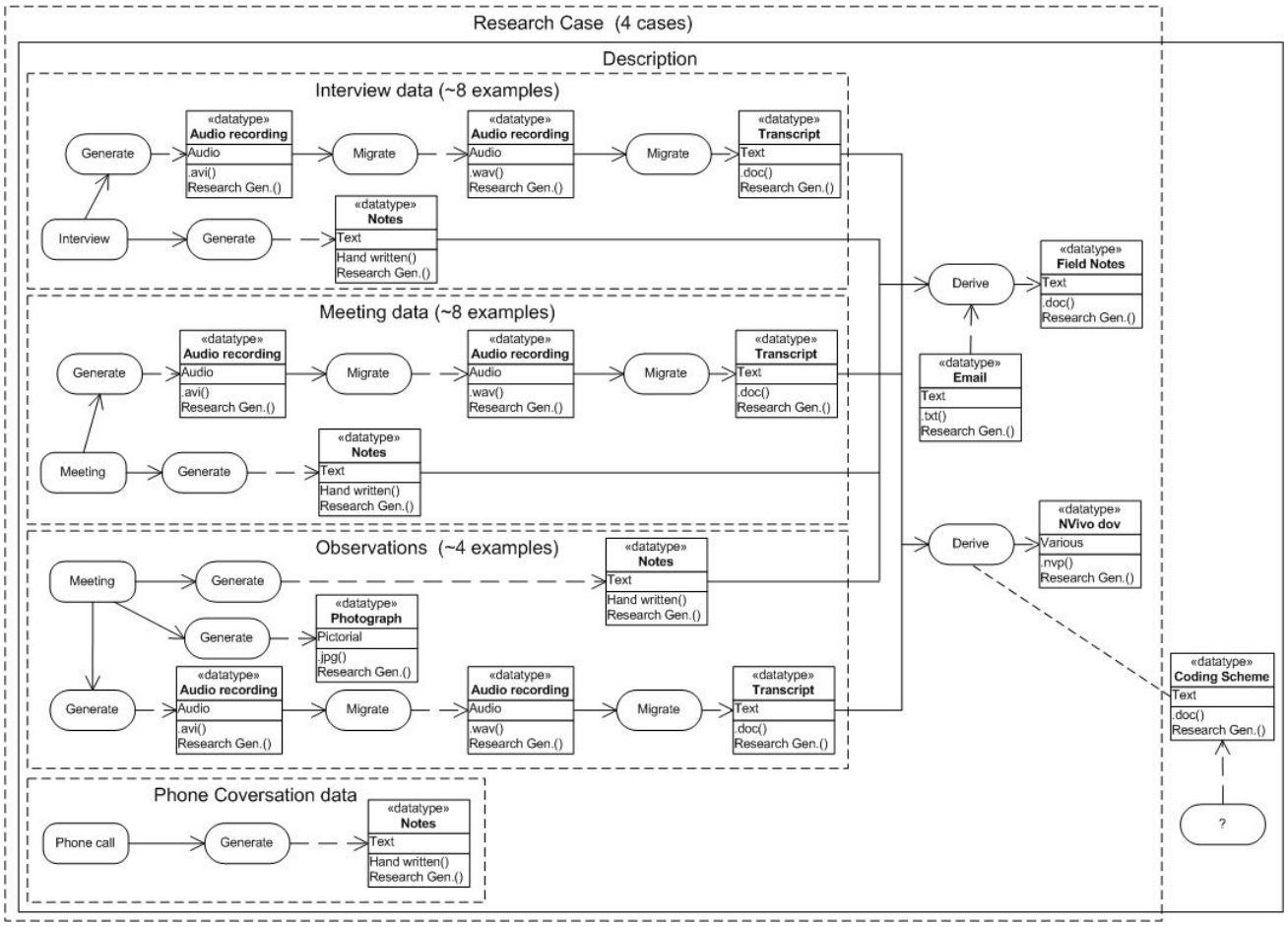

Figure 6-5. Understanding learning organisations RAID

6.6 Case Study 6: Cryogenic machining

This study showed two distinct RAIDs. The setup and parameterisation of both test rigs were described in the thesis but not maintained elsewhere. There are no links from the research records to the details regarding the setup or parameterisation. These RAIDs would be considered to be typical engineering descriptive studies, mapping the temperature profiles of a material and the cutting profiles.

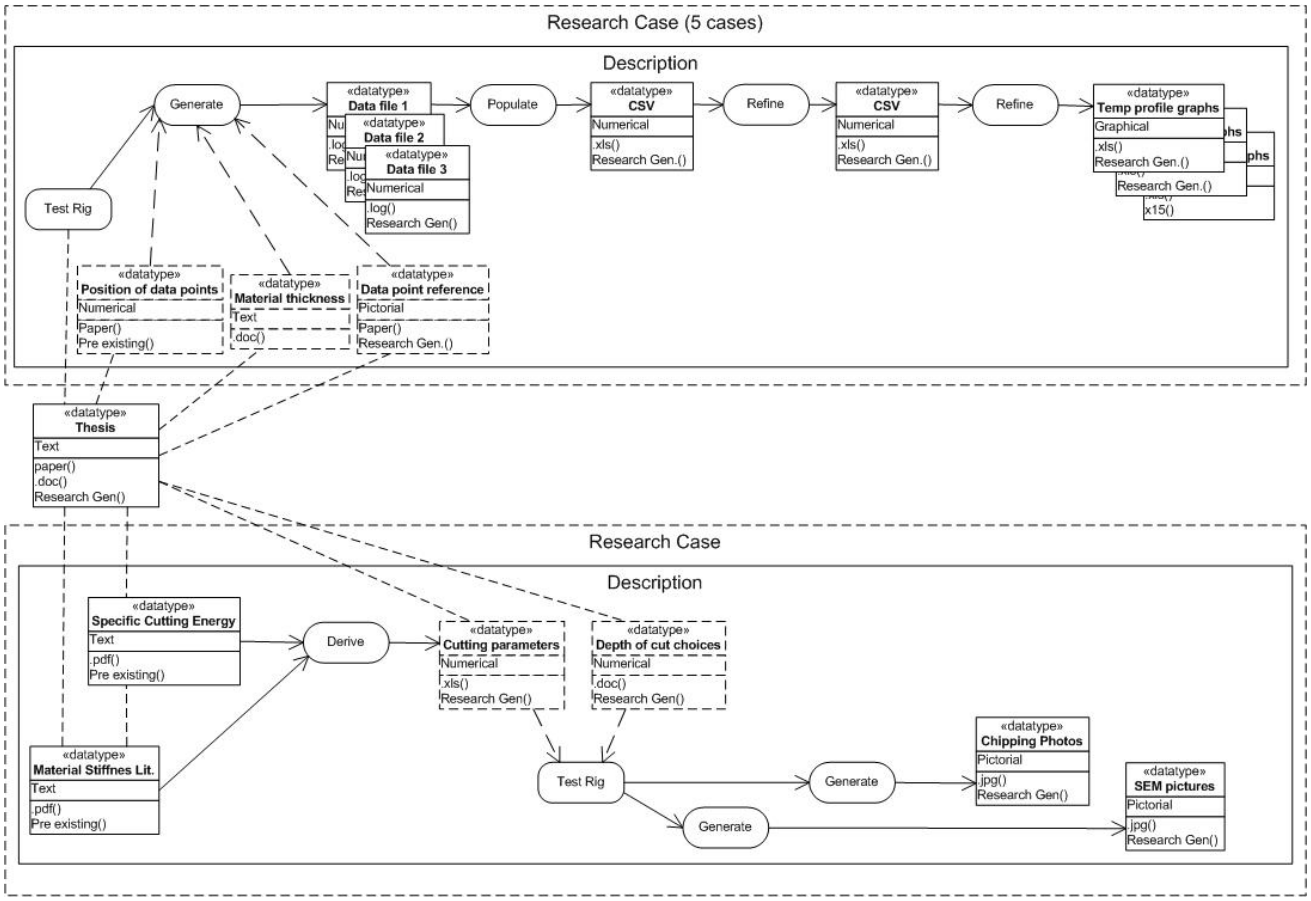

Figure 6-6. Cryogenic machining RAID

7. DISCUSSION OF RESULTS

The early part of this paper presented theoretical considerations relating to research activity and the development of research data. The data case audits provided the opportunity to validate some elements of the theory. The findings are presented here.

7.1 Validation of Terminology

Introduced in Table 3-1 was a set of data development processes to which research data is subjected during the research activity. Table 7-1 identifies where instances of these processes were found from the data case audits. Brief commentary on each process then follows.

Table 7-1. Data Development process validation based on instances of occurrence of each development process found in the data case audits

Data Development process	***Case number***					
	1	***2***	***3***	***4***	***5***	***6***
Addition	1					
Association						
Aggregation	1					
Annotation			1			
Augmentation						
Collection	1			2		
Collation						
Deletion						5
Derivation	3	2	2	7	2	1
Duplication						
Extraction			1			
Generation	1	5	2	2	8	3
Migration				2	6	
Population	1	2	1	1		1
Refinement				1		2

7.1.1 Addition

The one instance of adding data essentially details the further population of an already populated data record. In terms of the management, this may have some implication with regards to versioning but there should be little information loss.

7.1.2 Association

This was not found in any of the RAIDs yet should be considered vital in terms of information management. It was noticed that many of the research data assets within

the RAIDs were related (see dotted lines in Case Study 6) but crucially there was no association (that is to say, no explicit record of linking) between the documents. The authors believe that data record association is a foundational element in good data management; it is discussed further in Section 7.2.

7.1.3 Aggregation

One very clear instance of aggregation was identified. This was in the combining together of a number of different yet similar data assets – specifically, putting a set of questionnaire responses into a database. This has a number of important side effects and care should always be taken so that the data from the original records can be reproduced reliably. If entering into a spreadsheet, column and row manipulation can lose record of which data relate to which data asset/entry. At the same time intra-data information (for example ordering or clustering in raw data) can be lost.

7.1.4 Annotation

A single instance of annotation was found. Annotation is a very powerful means of providing addition information about the data itself, its context and so on as well as a means of describing data records. In particular it is one of the principal means of asserting/recording 'association'. The minimal occurrence of annotation (and, indeed, association) found in the data audit cases is not so much evidence of its uselessness, as evidence of the limited attention given to research management during the research process.

7.1.5 Augmentation

There was no evidence of augmentation. However, it is the nature of augmentation that it occurs without evidence remaining after the event, and therefore it would not necessarily become apparent in a post-event analysis as portrayed in a RAID diagram. It is quite possible that some of the data assets collected in Case Study 1 were augmented.

7.1.6 Collection

There were several instances of collection. Amongst other management considerations is whether the collected data can be referenced to the source or whether another copy must be kept and maintained.

7.1.7 Collation

No instances were found for collation. However, like augmentation this process may not be easily identified by a post-project analysis. Information loss is a likely accompaniment of this process; for example because changes in ordering occur during collation, information about earlier order may be lost, as too may be information about the source of the individual items of data.

7.1.8 Deletion

The act of deletion is unlike the other acts carried out in that it does not link two data assets. In Case Study 6 we can see five examples of temporary data assets that have at some stage been deleted (data assets with dashed line borders). We feel that understanding the impact of deletion of data from a set might demand a research

program on its own; it is clear that deletion has a profound impact on research information management. Deletion is defined as expunging or obliterating; there are, however, two similar acts which are related:

Disregarding: To look and ignore for the current purpose because the content is not important. The value then placed on the data may mean that subsequently it is managed in a way that is not conducive to data re-use or re-purposing.

Discarding: To use data or information then disregard it further.

Many issues surround the above. Policies on these forms of deletion should consider:

- Whether the information is reconstructable from the information kept
- Whether the deletion of the information makes other information non-reconstructable
- Whether there is a loss of context
- Whether the information will be useful (possibly for another purpose)
- Whether the information describes a one-off event
- Whether the information is kept elsewhere

All of the above considerations must be taken into account from the perspective of the value associated with the information. The evaluation information is outwith the scope of this research but is a substantial research subject in its own right.

7.1.9 Derivation

Instances of derivation were found in all cases. These instances can range from the highly objective and repeatable to subjective interpretations. There are many and varied information management implications, but it is fair to say that derived data, being inference or commentary on existing data, can be quite different from the original data. Deletion of data upon which a derivation is made is not a usual event; however, it is important that an association be made between source data and the data that are derived.

7.1.10 Duplication

Duplication was not captured in any of the RAID diagrams. However, if the questionnaires associated with the data cases are inspected it can be seen that the question templates would be *generated*; these would then be *duplicated* when sent to participants. The participant would *generate* and *add* the answers to the templates and send them back. The researcher would then *collect* and *aggregate* the results. It is known that duplication can cause all kinds of problems with version control. In this case it causes problems regarding unique identification of results. If a file with a particular name is sent to multiple participants, it can be expected to come back with the same name. Unless dealt with very carefully this process can have serious management implications. Online surveys, suitably conducted, can eliminate many of the negative side effects.

7.1.11 Extraction

This is essentially the reverse of aggregation. A specific part of a file is removed for further analysis or independent work. This was identified in Case 3 where individual diagrams from a set were extracted into another package which had additional functionality.

7.1.12 Generation

This is perhaps the core activity in engineering research: creating data about one's subject. This development process has the most management implications. As there is no preceding data in the RAID there is little context into which to put the data asset. All of the contextual information about the generated data asset must be captured within the asset or referred to in a contextualising document. During this process lots of data is both *discarded* and *disregarded*.

7.1.13 Migration

There were many instances of this. In a research context format migration is often to put the data into a format that enables extra functionality. For example, migrating information from a proprietary format to a standard format which can be manipulated in another software package (identified in Data Cases 4 and 5).

7.1.14 Population

This is commonplace in engineering research. It is not a surprise that the only case that did not contain this development process also contained no prescriptive study. In most cases the prescriptions were some form of modelling software which needed to be populated with information.

7.1.15 Refinement

Rather surprisingly only three instances of refinement could be identified from the data cases. Nevertheless, refinement does have management implications because the side-effects include information loss at the data level. For good management, it may be necessary for precursor data to be retained or for contextualizing information to be retained which allows – where it is possible – the source data to be recovered, either by revisiting the source or by reversing the refinement process.

7.2 Association of Data Assets

Many of the data assets identified in the RIAD diagrams were related in the sense they had shared topics or information but none were explicitly associated. At the same time many of the assets represented implicitly the context in which other records could be interpreted. At present the means for associating data assets is laborious and cumbersome. Good practice would involve bibliographical referencing and inventory lists or audits, or the annotation of metadata either in the record itself or in stand-alone documentation.

Likewise, hyper-linking to the source can be very effective for associating data assets. As demonstrated by the dearth of such actions in the data audits, however, explicit association of data is not a defining characteristic of current research data management (such as there is). Impediments to making explicit associations include technical and cultural ones, not least that researchers are neither given (nor expected to ask for) the resources and tools necessary for such data management to take place. The authors feel strongly that explicit association of the data assets in a research activity is the foundation of the sort of good data management that would lead to easier data re-use. To make such associations practicable and to limit the involvement of the researcher in making them the authors feel that as much reliance should be made as possible on

automated processes, rather than relying on additional work by the researcher, whose efforts are better directed at the research itself.

Creating automatic associations between data assets is technically feasible and would provide the context necessary to make the data assets more findable, interpretable, verifiable, repeatable, replicable and useful. The RAID modelling method could be employed as the basis for tracking and recording associations, as well as being the basis for visualizing the 'map' of the research activity information development.

The means for associating data or data records can be grouped into two main classes as shown in Figure 7-1. In the first class the association data is carried locally with the data or data record; in the second class the information about the association is carried in stand-alone documentation remote from the data that are being associated. An example of the first class is where metadata is included in the data record itself, referring either to the data within or to the data record. These data might be recorded in a properties box or alternatively in, say, an XML encoding.

Across these two classes there are potentially seven different strategies for associating data assets. The arrows indicate the subject of the association and where the association data is carried. So for example in Strategy 1, the two assets are linked by metadata or annotation to be found in Data Asset 1, with no reciprocal association data in Data Asset 2. Strategy 2, where the association information is reciprocal, might be the preferred arrangement. All five strategies in the right-hand partition in Figure 7-1 are associated through a stand-alone document. It can be seen, however, that for complete association to occur a full set of reciprocal references are necessary, as indicated in Strategy 7. Clearly, this being the most complete explicit association, it would be the preferred one in principle for good management.

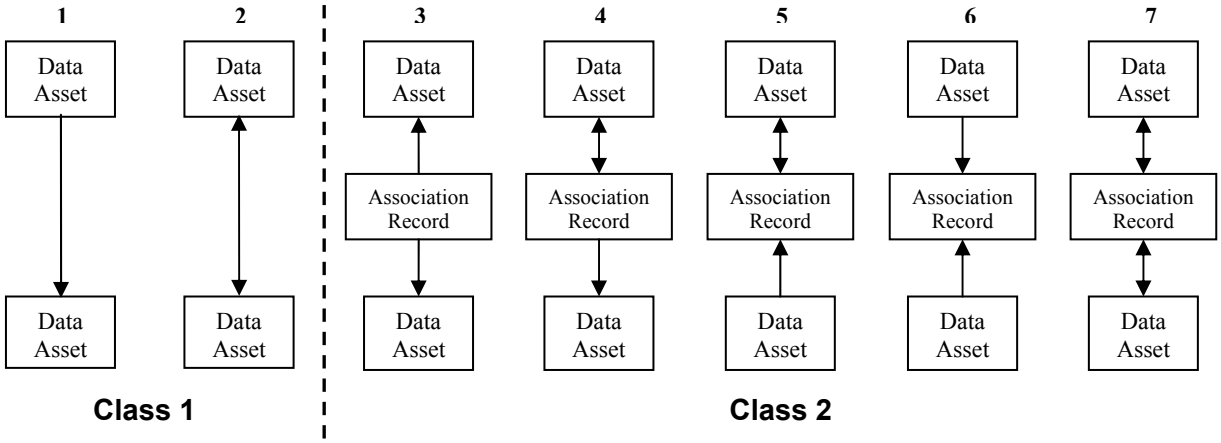

Figure 7-1. Strategies 1-7 for associating data assets

It should be recognized that these strategies of association could be used to record the association of research data records with more abstract things such as tags, labels, cases, projects and so on, thus providing rich contextual information. Also, associations are not limited to one-to-one relations, but can extend to one- or many-to-many.

8. CONCLUSION

The research described in this report has contributed considerably to the greater understanding of the diversity and character of engineering research data and provided the basis for improved methods of its management for re-purposing and re-use. Part of

the work has been development of an emerging terminology to help describe the different types of engineering information and the different forms of development process and management activity to which it is subjected to during research. In particular the notions of ‘data purposing’, ‘data re-purposing’ and ‘supporting data re-use’ have been identified as data preparation activities which motivate research data management, in the latter two cases for re-use.

The concepts defined in the terminology provide the elements in a new means of modelling the research activity and the development of associated research and contextual data. These data are recorded in a number of types of data record identified here by the authors. This theoretical work is informed and to some extent validated by the findings of, first, a scoping survey and, then, an audit of selected cases of engineering research data. Scrutiny of the data, and investigation of the research by which it was gained, provides the basis for characterizing a broad spectrum of data, the development of which is captured for each audited data case, using the new modelling method, in a Research Activity Information Development (RAID) diagram. A distinction is drawn in characterizing research data between that associated with ‘descriptive’ research on the one hand and ‘prescriptive’ research on the other, the second type being very common in engineering research work. It can be observed, however, that because of its nature and specificity, re-use and repurposing of prescriptive research data may be inherently more difficult that that of descriptive research.

In considering the different types of development process to which research data is commonly subjected, the authors believe that ‘association’ is perhaps foundational in supporting good data management for its easier re-use and re-purposing. To support the researcher in the better management of data as it is developed, and to provide a ‘map’ to aid its later understanding, the authors propose the use of an automated ‘association’ tool, based on the information that can be captured and represented in a RAID diagram. By capturing and recording the development of research data in individual activities or projects it is proposed that data assets can be made more findable, interpretable, verifiable, repeatable, replicable and useful.

9. REFERENCES

Ball, A. (2010). *Review of the state of the art of the digital curation of research data* (ERIM Project research report erim1rep091103ab12). Retrieved September 15, 2010 from http://opus.bath.ac.uk/19022

Beagrie, N. & Jones, M. (2001). *Preservation management of digital materials: A handbook.* London: British Library. ISBN: 0-7123-0886-5. Retrieved September 15, 2010 from http://www.dpconline.org/advice/digital-preservation-handbook.html

Beagrie, N., Beagrie, R., & Rowlands, I. (2009). Research data preservation and access: The views of researchers. *Ariadne, 60*. Retrieved September 15, 2010 from http://www.ariadne.ac.uk/issue60/beagrie-et-al/

Blessing, L. & Chakrabarti, A. (2009). *DRM: A design research methodology.* London: Springer. ISBN-13: 978-1-84882-586-4.

Birnholtz, J. & Beitz, M. (2003). *Data at work: Supporting sharing in science and engineering.* In M. Pendergast (Ed.), *Proceedings of the 2003 International ACM SIGGROUP conference on supporting group work* (pp. 339–348). New York: Association for Computing Machinery. ISBN: 1-58113-693-5. DOI: 10.1145/958160.958215.

Burton, A. & Treloar, A. (2009). Designing for discovery and re-use: The 'ANDS Data Sharing Verbs' approach to service decomposition. *International Journal of Digital Curation, 4*(3), 44–56. Retrieved September 15, 2010 from http://www.ijdc.net/ijdc/article/view/133

Cohen, F.S. (1950). Field theory and judicial logic. *Yale Law Journal, 59*, 238–272.

Consultative Committee for Space Data Systems. (2002). *Reference model for an Open Archival Information System (OAIS)* (Blue Book CCSDS 650.0-B-1). Also published as ISO 14721:2003. Retrieved September 15, 2010 from http://public.ccsds. org/publications/archive/650x0b1.pdf.

Darlington, M.J., Culley, S.J., Zhao, S., Austin, S.A., & Tang, L.C.M. (2008). Defining a framework for the evaluation of information. *International Journal of Information Quality, 2*(2), 115–132. Retrieved September 15, 2010 from http://hdl.handle.net/2134/4226

Darlington, M., Heisig, P., Leiringer, R. & Burt, G. (2009). *KIM Resources Inventory* (KIM Project document kim50too007gb10).

Digital Curation Centre. (2007, April 26). *What is digital curation?* Retrieved September 15, 2010 from http://www.dcc.ac.uk/digital-curation/what-digital-curation

Evrard, F. & Virbel, J. (1996). *Realisation d'un prototype de station de lecture active et utilsation en milieu professionnel* (Rapport du contrat 9300571). Toulouse: ENSEEIHT INPT.

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J., & Rasmussen, B. (2008). *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes* (Final report). London: JISC. Retrieved September 15, 2010 from http://ie-repository.jisc.ac.uk/279/

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation, 3*(1), 134–140. Retrieved September 15, 2010 from http://www.ijdc.net/ijdc/article/view/69

Jones, S., Ball, A. & Ekmekcioglu, C. (2008). The Data Audit Framework: A first step in the data management challenge. *International Journal of Digital Curation, 3*(2), 112-120. Retrieved September 15, 2010 from http://www.ijdc.net/ijdc/article/view/91

Jones, S. (2009). *A report on the range of policies required for and related to digital curation* (Deliverable H1.1). Version 1.2. Edinburgh: Digital Curation Centre. Retrieved September 15, 2010 from http://www.dcc.ac.uk/sites/default/files/documents/reports/DCC_Curation_Policies_Report.pdf

KIM Project. (2009). *Knowledge and Information Management (KIM) Grand Challenge Project.* Retrieved September 15, 2010 from http://www.kimproject.org/

Lord, P. & Macdonald, A. (2003). *Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision* (E-Science Curation Report). London: JISC. Retrieved September 15, 2010 from http://www.jisc.ac.uk/media/documents/programmes/preservation/e-sciencereportfinal.pdf

Van Beveren, J. (2002). A model of knowledge acquisition that refocuses knowledge management. *Journal of Knowledge Management, 6*(1), 18–22.

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a standard for the social sciences. *International Journal of Digital Curation, 3*(1), 107–113. Retrieved September 15, 2010 from http://www.ijdc.net/ijdc/article/view/66

Williams, R. & Pryor, G. (2009). *Patterns of information use and exchange: Case studies of researchers in the life sciences* (RIN report). London: Research Information Network & British Library. Retrieved September 15, 2010 from http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/disciplinary-case-studies-life-sciences

10. APPENDIX A. DATA MANAGEMENT TERMINOLOGY V3

Notes

This terminology is the subject of continuous development during the project, partly as a result of further thinking by the ERIM team, and partly as a result (we hope) of contributions from the JISC Data Management community. The terminology recorded here represents the terminology state at the date of issue of this document.

The terms are not strictly in alphabetical order but have been grouped loosely into logically associated clusters; the clusters are demarked by horizontal lines.

The terms and their definitions have come from a number of sources. Where possible the provenance of a term that has been borrowed is given against the entry.

The 'data discovery verbs' coined and defined by the Australian National Data Service (ANDS) are not included here. This is partly because a number of their definitions conflict with those here, and partly because their verbs are task-specific (providing a structure and high-level architectural device to support services necessary to sharing) and our task (managing data for the purposes of re-use and re-purposing) is somewhat different and more diverse.

Amongst the terms is a set of words defined which can be applied in noun form (when related to an act or process being carried out on data – for example association or augmentation) or as a verb (for example, associate or augment). The verb form is used in the terminology (for alignment with the ANDS verbs). An understanding of the meaning of both forms of the word should be apparent from the accompanying definition.

Data Reinterpretable representations of information in a formalized manner suitable for communication, interpretation or processing.

Examples of Data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen. (OAIS Reference Model)

Information Any type of Knowledge that can be exchanged. In an exchange, it is represented by Data. An example is a string of bits (Data) accompanied by a description of how to interpret a string of bits as numbers representing temperature observations measured in degrees Celsius (Information). (after OAIS Reference Model)

Note 1. The 'accompanying description' referred to above could be substituted by Knowledge appropriate to achieving the same interpretation.

Note 2. There is a logical tension apparent in the definitions of Data and Information in respect of the enterprise of making 'research data' re-usable. Research Data are only re-usable if the Information or Knowledge is (made) available for correct interpretation to follow, in which case strictly it is 'Information' not 'Data'.

Knowledge The stock of Information, skills, experience, beliefs, and memories existing in the head of the individual. (after Van Beveren, 2002)

Metadata Data about Data (OAIS Reference Model)

The term 'metadata' is often used in a rather restricted way, for example limiting the scope to keywords, terms populating (XML, HTML) tags and so on. It is intended here to be interpreted to embrace 'Information about Data' , and includes both Data that is local to the Data that it describes (i.e. in the same Data Object) or is held remotely in a separate file. All Metadata 'contextualizes' the Data to which it refers; some Metadata is a means by which Data is associated

---

Research Activity The process through which research Data and context Data are accumulated and developed.

Research Data Development Process One of a set of processes that are commonly carried out during the Research Activity which changes or adds to the research Data associated with a Research Activity or project.

Data Case The set of Data Records associated with some discrete Research Activity (project, task, experiment, etc.).

Record Information in any medium, created, received and maintained as evidence of an activity.

Associative Data Record (ADR) A Record which documents the association between other Data or Data Records. The Data contained within an ADR is a special case of contextualizing Data and the Data Record a special case of a CDR.

Data Record (DR) The Data Object which contains the Data.

Experimental Apparatus Data Record (EADR) A Digital Object which is analogous to the physical experimental apparatus familiar in much laboratory-based research.

Research Data Record (RDR) A Record containing research Data, i.e. Data that is descriptive of the research object.

Context Data Record (CDR) A Record containing Data explicitly intended to place in context other Data or abstract aspects of the Research Activity or subject.

A special case of the CDR is the Associative Data Record.

Research Object Data Record (RODR) A Data Object which is itself the object of research interest or which together with other RODRs constitutes the object of research interest.

Data Object Either a Physical Object or a Digital Object. (OAIS)

Digital Object An object composed of a set of bit sequences. (OAIS)

Manifestation The way in which the intangible underlying Data (contained in a file) is embodied (i.e. presented for interpretation).

Note. Underlying Data can be manifest in different ways; e.g. as a spreadsheet, as a graph, or as an HTML or PDF page.

Add To supplement existing Data at the Data level, for example when inducting Data into an existing file.

Aggregate To combine similar Data from different sources for the purpose of increasing sample size (cf. Augment).

Annotate To add 'information or additional marks formulated on a document for enhancing it with brief and useful explanations' (Evrard & Virbel, 1996).

Note. Annotation may be made on the subject document itself or on a 'stand-off' document with the annotation linked to the annotated content.

Associate To make explicit the relationship between items of Data, Data Records or Data Cases (cf. Related).

Example. Two Data Records may be related by such things as file names, metadata, explicit reference in one to another, by an embedded link or by a separate document which cites the Association (see: Associative Data Record).

Augment To add Data Records to a Data Case.

Collate Give order to Data assembled from different sources.

Note. This can occur at any organizational level, that is at the Data level, the Data Record level or the Data Case level.

Collect To acquire and bring together pre-existing Data.

Note. This is concordant with the DCC's use of the term 'receive' in relation to pre-existing data collected from external sources.

Delete To expunge or obliterate.

Pretty self-explanatory, but rather important in relation to data management.

Derive To create new Data by applying logical inference, extrapolation, or similar algorithmic process to pre-existing Data (cf. Refine).

Note 1. Derivation constitutes creating a new description or representation of an analysis of the subject described by the existing Data.

Examples. A histogram representing the frequencies of Data occurrence; a narrative or commentary on a set of interviews.

Note 2. NASA's data-processing terminology includes the use of the term 'derivation' with a similar interpretation; however it seems to conflate the three concepts of 'derivation' and 'refinement' and 'manifestation'. There 'derived data' are defined as 'derived results, as maps, reports, graphs, etc.'; this being data processed through 'NASA Process Levels 2 though 5'. There is some argument to say that 'derivation' and 'refinement' occupy different positions on the same continuum. However, they seem to be different in character and thus may have

different management implications. If it is found that they do not the two concepts could usefully be conflated (as in the NASA definition).

Duplicate To make another, identical, copy of a file.

Extract To make a new Research Data Record from portions of the Data in one or more Research Data Records.

Generate To act on or interact with a research subject thereby creating research Data.

Migrate (Format Migration) The transfer of digital information from one format to another (with the intention of preservation of the full information content) (DCC glossary)

Populate To enter Data residing in a Data Record into an existing framework such as a pro-forma or a Knowledge- or Data-base.

Refine To re-express Data in a different form or according to a different data model (cf. Derive).

Example of typical refinement functions: rounding, normalization, removing duplicates, stop-words, noise or outliers, simplification, etc.

Clean (Data) A special case of Refinement, where the refined Data are a normalized version of the source Data; that is, with systematic errors corrected, calibration taken into account, invalid Data removed, etc.

Record (Data) Encoding the output from Data Generation in a carrier format (e.g. bitstream, written notes).

Transform To derive or refine Data in such a way that new or changed Data results.

Note. this interpretation excludes Format Migration because in Migration – in contrast to derivation and refinement – there is expressly no intention to change the Data.

First-Generation Data The Data resulting either from Collection or from Generation.

This term is intended to identify Data which has not been the subject of Derivation or Refinement

Data Rawness The inverse measure of the number of processing steps leading to the creation of a set of Data.

Information Loss Removal of Information from an instance or set of Data.

Examples are such things as rounding down or up of real numbers to integers, Deletion of the record of units in a Data set, or disassociation of context Data and the Data it explains.

Information Gain Addition of Information to an instance or set of Data, for example when Aggregating or Annotating.

Function loss Removal of or reduction in the capacity to compute or manipulate.

An example is migrating the contents of a live spreadsheet to a PDF format where the content stays the same, but the facility changes.

Function Gain Increase in the support for computation or manipulation of Data

An example is transferring the Data in a hand-written sheet into a spreadsheet providing such facilities as ordering, summing, etc. Likewise, Function Gain is a characteristic of format change through optical character recognition.

State Loss Erasure or discarding of earlier Data state(s) or version(s).

State Loss occurs either as a result of discarding unregarded Data or overwriting an existing version or the Data. It characteristically occurs when the item of interest is the final outcome of an iterative process, for example in the application of the Delphi Method or when continuous updating of Data occurs, for example in the automatic updating of computer code in response to closed-loop feedback.

Related Two or more items of Data, or Data Records which have an implicit or explicit connection. Explicit connections are made through Association.

Process Repeatability A measure of the practical possibility of a Data-gathering process being repeated such that Data Reproducibility is possible in principle.

Data Reproducibility Data is reproducible if it can be regenerated through repeat of a Data-gathering process such that its functional content remains the same.

Knowing the in-principle ability or non-ability to reproduce (or re-gather) Data is important for Data management; it impinges not only on considerations of Data acquisition and maintenance but also experimental repeatability, inference validity and so on. However, the interpretation of repeatable is legitimately variable dependent on process and requirement. In some disciplines strict repeatability is a requirement for experimental and inferential validity, in others the concept is a non sequitur. In some processes, identical input will produce identical output. Some processes, given identical input, will produce the same average output, from which the same conclusions can be drawn. For some processes the concept of identical input and output are inappropriate concepts; what will be of interest is whether for the same general input conditions the same interpretation or inference can be drawn.

Reversible True of a process P if and only if there can exist, at least in principle, a process P′ which, when given P(I) as input, produces I as output.

Non-Reversible True of a process P if and only if there cannot exist, even in principle, a process P′ which, when given P(I) as input, produces I as output.

Data Use Using research Data for the current Research Activity or purpose to infer new Knowledge about the research subject.

Data Re-use Using research Data for a Research Activity or purpose other than that for which it was intended.

Supporting Data Re-use Managing existing research Data such that it will be available for a future *unknown* Research Activity.

This concept is one for which no verb has been coined. It is one of a set which also includes Data 'Purposing' and 'Re-purposing'. It combines many of the activities implied in the verbs 'archive', 'preserve' and 'curate'.

Data Purposing Making research Data available and fit for the current Research Activity.

This is the activity all researchers are familiar with when making research Data available for their own research.

Data Re-purposing Making existing research Data available and fit for a future *known* Research Activity.

Note. This definition emphasizes the activity as being one of explicit intention, and thus differs (as does the spelling) from the definition used for 'repurposing' in the Data Documentation Initiative's Combined Life Cycle Model for research data, viz:

> Repurposing. *The data may also be used within a different conceptual framework; examples include sampling or restructuring the data, combining the data with other similar sets, or producing pedagogic materials.*

11. APPENDIX B. SCOPING SURVEY QUESTIONNAIRE.

This questionnaire was used in order to gain a first understanding of the diversity and character of engineering research data.

Q1. What media type is this data?
Textual
Video
Pictorial (2D)
Pictorial (3D)
Audio
...
Q2. Is this data of a real of simulated situation?
Reality (real industrial observations)
Simulation (e.g. role playing students)
Q3. What file format if the data saved as?
Physical
Electronic:
- Word doc
- Excel
- Power point
- …

Q4. Is this data derived from any other data?
– Were there any steps between your answer and the data?
1st-Generation
Refinement
Derivation
Q5. Where did the original data come from?
Pre-existing
Research-generated
Q6. What method did you use to collect the original data?
Data from surveying through questionnaires and interviews
Data from transcribing discussions
Data from experimental rigs
Data from commercial software
Data from a bespoke version of commercial software
Data from fully bespoke software
Data from discursive process
Q7. How repeatable are these results/data?
Repeatable
Non-repeatable
Q8. How repeatable was the data creation process?
Repeatable
Non-repeatable
Q9. What type of interpretation is likely to be or was applied to this data?
Objective
Subjective
Q10. How do you expect this data record to change over time?
Open (ternary)

Expanding - being added to
Dynamic - being modified
Expanding and Dynamic
Closed
Definitive
Dormant

Q7. How repeatable are these results/data?

Repeatable
Non-repeatable

In light of some of the future work packages 3 other questions were also added:

Q11. What is the sample size and units?
Q12. What are the barriers to the re-use of this data?
Q13. Any further notes/comments?

12. APPENDIX C: A DIGRESSION RELATING TO THE TRICKY QUESTION OF IDENTIFYING SORTS OF SYMBOLIC REPRESENTATION

In Section 3.2.4 a discussion occurred which contrasted 'conventional' research activities – which employ both physical and data-based artefacts – with research in which the artefacts consist in symbolic representation alone. In the second type of research it can be difficult to identify the rôle played by particular data assets or records and, indeed, the rôle of any given data may not only change during the course of the research, but it can fulfil a number of different rôles depending on the perspective from which the data is viewed.

By way of illumination of such difficulties of identification, two research examples will be contrasted; a 'conventional' research experiment and a 'paper'-only-based activity, of the sort commonly undertaken in information and knowledge management research and in computational-based research.

The conventional research activity chosen is the development and testing of a method of transmuting Material A into Material B by carrying out a process by means of an experimental rig (as illustrated in the upper element of Figure 3-4). The research data in this activity might consist of a description and values associated with Materials A and B, a description of the laboratory method of transmutation and of the experimental laboratory rig, and a written report and analysis of the outcome. It is quite clear in this research that there are, on the one hand, data, and on the other the physical components involved, these being the materials and the laboratory rig. The data is referred to here as being contained in one or more data records including context data records (CDRs), Research Data Records (RDRs) and Associative Data Records (ADRs).

Now take another research activity where research is being carried out into finding a means of recording and representing the design task in a more desirable way than possible hitherto. In this research there exists an event (the design episode) which is recorded in some manner, a method of transforming the raw record of the event into one that fulfils the criteria being sought, and descriptions of the input, the output, the transmuting function and so on. The transmuting function may well consist of computational code, together with initialization values and a specification of the expected inputs and outputs to it.

In the first case it is clear where exists the data and where the physical elements of the experiment, parts of which constitutes the object of the enquiry. In the second case it would not be clear where data management started and stopped. Many real research projects are considerably more complex than this illustration, where symbolic representations are often nested, and play different rôles at different times in the research, sometimes being 'data', sometimes entities analogous to experimental apparatus, sometimes the object of research itself, sometimes explicitly contextualizing data, and so on.

In the second case, the data and the representational elements of the object under scrutiny are indistinguishable since both are symbolic representations. In management terms, the question arises as to which parts of the representational material should be subjected to management. Clearly, the representational analogues of the physical 'experiment apparatus' could be construed as 'contextual data' according to the definition, but should they? And how should the object of research be treated? To

exemplify further, take an instance of a questionnaire designed to elicit information, and a data sheet which encodes the responses to this questionnaire by each of the individuals in a set of participants. It is clear how the results data should be classed (research data), but does the questionnaire represent an analogue of physical experimental apparatus or is it 'contextual' data?

Understanding the differences between research data proper, context data and 'stuff that is the paraphernalia of research', that is, analogous to experimental apparatus, must be important in understanding research data development, information flow, and the management implications. At this stage the understanding has not become clear. However, to achieve some clarity (where different symbolic records can be found which are analogous to physical elements of an experiment) the terms Experimental Apparatus Data Record (EADR) and Research Object Data Record (RODR) are used which, if nothing else, allows labels to be used and the data assets of this nature to be referred to.